

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

FABRÍCIO CARVALHO MACIEIRA

ALÉM DA FISCALIZAÇÃO TRADICIONAL: UTILIZANDO INTELIGÊNCIA
ARTIFICIAL PARA SELEÇÃO DE AMOSTRA DE AUDITORIA

SÃO PAULO

2025

FABRÍCIO CARVALHO MACIEIRA

**ALÉM DA FISCALIZAÇÃO TRADICIONAL: UTILIZANDO INTELIGÊNCIA
ARTIFICIAL PARA SELEÇÃO DE AMOSTRA DE AUDITORIA**

Dissertação apresentada à Escola de Administração de Empresas de São Paulo da Fundação Getúlio Vargas, como requisito para obtenção do título de Mestre em Gestão e Políticas Públicas.

Linha de Pesquisa: Políticas Públicas Urbanas

Orientador: Prof. Dr. Joelson Sampaio

SÃO PAULO

2025

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Macieira, Fabrício Carvalho.

Além da fiscalização tradicional: utilizando inteligência artificial para seleção de amostra de auditoria / Fabrício Carvalho Macieira. - 2025.
69 f.

Orientador: Joelson Oliveira Sampaio.

Dissertação (mestrado profissional MPGPP) – Fundação Getulio Vargas, Escola de Administração de Empresas de São Paulo.

1. Administração pública - Inovações tecnológicas. 2. Inteligência artificial. 3. Aprendizado do computador. 4. Empresas públicas - Auditoria. I. Sampaio, Joelson Oliveira. II. Dissertação (mestrado profissional MPGPP) - Escola de Administração de Empresas de São Paulo. III. Fundação Getulio Vargas. IV. Título.

CDU 35.073.526

Ficha Catalográfica elaborada por: Isabele Oliveira dos Santos Garcia CRB
SP-010191/O

FABRÍCIO CARVALHO MACIEIRA

**ALÉM DA FISCALIZAÇÃO TRADICIONAL: UTILIZANDO INTELIGÊNCIA
ARTIFICIAL PARA SELEÇÃO DE AMOSTRA DE AUDITORIA**

Dissertação apresentada à Escola de Administração de Empresas de São Paulo da Fundação Getúlio Vargas (FGV EAESP), como requisito para obtenção do título de Mestre em Políticas Públicas.

Linha de Pesquisa: Políticas Públicas Urbanas

Data de avaliação: 23/05/2025

Banca examinadora:

Prof. Dr. Joelson Sampaio

Orientador FGV-EAESP

Prof. Dr. Marco Antonio Carvalho Teixeira

Professor FGV-EAESP

Prof. Dr. Caio César de Medeiros

Universidade de Brasília (UNB)

AGRADECIMENTOS

Agradeço, antes de tudo, à minha companheira Paula, por toda compreensão e incentivos nesses longos meses de pesquisa.

Aos meus pais, que sempre apoiaram meus estudos e me motivaram a chegar aqui.

Ao Tribunal de Contas do Estado de São Paulo, por proporcionar essa experiência tão enriquecedora, espero poder retribuir à altura.

Por fim, mas não menos importante, ao meu orientador, Professor Joelson, por todo apoio e dedicação.

“I may not have gone where I intended to go, but I think I have ended up where I needed to be” — Douglas Adams

RESUMO

O avanço da Inteligência Artificial (IA) tem transformado significativamente a forma como organizações públicas e privadas realizam auditorias, possibilitando a análise de grandes volumes de dados e aumentando a capacidade de identificar riscos com maior precisão e agilidade. No setor público brasileiro, esse avanço representa uma oportunidade para otimizar o uso de recursos humanos diante do desafio de fiscalizar um grande volume de recursos. Esta pesquisa teve como objetivo aprimorar um modelo preditivo baseado em algoritmos de *machine learning* para auxiliar a seleção de amostras de auditoria no Tribunal de Contas do Estado de São Paulo (TCE-SP), desenvolvendo uma ferramenta intuitiva para uso dos auditores. A metodologia adotada consistiu em uma pesquisa aplicada, com abordagem quantitativa, utilizando dados públicos de execução orçamentária, dados demográficos, econômicos, educacionais e de saúde de 644 municípios paulistas. Três algoritmos supervisionados foram comparados: *Random Forest*, *Support Vector Machine* e *LightGBM*. O modelo *Random Forest* apresentou o melhor desempenho, sendo otimizado com ajuste de hiperparâmetros e validação cruzada. Os resultados demonstraram que a melhoria da qualidade das variáveis utilizadas e a otimização dos hiperparâmetros melhoraram a capacidade preditiva do modelo. Além disso, foi proposta uma metodologia prática para o uso dessas previsões na seleção de amostras de auditoria, baseada em limites de confiança empíricos. Como contribuição, o estudo oferece uma ferramenta acessível aos auditores, com potencial de aplicação em outros conjuntos de dados, que pode ser incorporada aos processos de planejamento de auditoria, aumentando a capacidade de análise das equipes responsáveis.

Palavras-chaves: Administração Pública, Auditoria Pública, Inteligência Artificial, Machine Learning, Random Forest, SVM, LightGBM.

ABSTRACT

The progress of Artificial Intelligence (AI) has significantly transformed the way public and private organizations conduct audits, expanding the possibilities of analyzing large volumes of data and increasing the ability to identify risks with greater precision and speed. In the Brazilian public sector, this advancement represents a strategic opportunity to optimize the use of human resources in face of the growing challenge of overseeing large volume of public funds. This research aimed to improve a predictive model based on machine learning algorithms to support the audit sample selection process at the São Paulo State Court of Accounts (TCE-SP), by developing an intuitive tool for auditors. The adopted methodology consisted of applied research with a quantitative approach, using public data on budget execution, as well as demographic, economic, educational, and health data from 644 municipalities in the state of São Paulo. Three supervised algorithms were compared: Random Forest, Support Vector Machine, and LightGBM. Random Forest presented the best performance and was further optimized through hyperparameter tuning and cross-validation. The results showed that improving the quality of the updated variables and optimizing the hyperparameters enhanced the model's predictive capability. Additionally, a practical methodology was proposed for using these predictions in audit sample selection, based on empirical confidence limits. As a contribution, the study offers an accessible tool for auditors, with potential for application in other datasets, that can be incorporated into audit planning processes, increasing the analytical capacity of audit teams.

Keywords: *Public Administration, Public Auditing, Artificial Intelligence, Machine Learning, Random Forest, SVM, LightGBM.*

LISTA DE FIGURAS

Figura 1. Modelo de RF esquematizado	21
Figura 2. Parâmetros da consulta à API Sidra	32
Figura 3. Boxplot das variáveis demográficas e estruturais (Exercício de 2023)	38
Figura 4. Boxplot das variáveis econômicas e fiscais (Exercício de 2023)	39
Figura 5. Boxplot das variáveis educacionais (Exercício de 2023).....	40
Figura 6. Boxplot das variáveis da saúde (Exercício de 2023)	41
Figura 7. Valor total da despesa liquidada em 2023 por tipo de entidade	42
Figura 8. Dez elementos que representam o maior percentual da despesa executada em 2023	43
Figura 9. Tempo de treinamento dos três modelos	45
Figura 10. Importância das variáveis no modelo RF otimizado.....	49
Figura 11. Tempo de treinamento do modelo RF otimizado	53
Figura 12. Gráfico comparativo com os valores previstos pelo modelo e aqueles efetivamente executados	54
Figura 13. Gráfico comparativo com os valores previstos pelo modelo e aqueles efetivamente executados	55
Figura 14. Resultado da otimização dos hiperparâmetros no modelo Random Forest	58
Figura 15. Resultado da otimização dos hiperparâmetros no modelo LightGBM	58

LISTA DE TABELAS

Tabela 1. Primeiros 10 desdobramentos do elemento 36 utilizados no exercício de 2023	16
Tabela 2. Variáveis utilizadas para treinamento dos modelos preditivos	30
Tabela 3. Amostra de valores corrigidos pelo IPCA até dezembro de 2023	33
Tabela 4. Estatísticas descritivas das variáveis utilizadas	35
Tabela 5. Subdivisões do elemento de Despesa 04 – Contratações por tempo determinado ...	43
Tabela 6. Avaliação dos modelos.....	45
Tabela 7. Melhores parâmetros encontrados para o modelo Random Forest.....	47
Tabela 8. Melhores parâmetros encontrados para o modelo LightGBM.....	47
Tabela 9. Comparação dos erros dos modelos após otimização (Random Forest).....	48
Tabela 10. Comparação dos erros dos modelos após otimização (LightGBM)	49
Tabela 11. Correlação entre as variáveis que representam a quantidade de matrículas	50
Tabela 12. Correlação entre as variáveis que representam a quantidade de docentes.....	50
Tabela 13. Correlação entre as variáveis que representam a quantidade leitos hospitalares	51
Tabela 14. Correlação entre as variáveis que representam o PIB municipal.....	51
Tabela 15. Comparação dos modelos padrão antes e após a exclusão das variáveis altamente correlacionadas	52
Tabela 16. Comparação dos modelos otimizados antes e após a exclusão das variáveis altamente correlacionadas	52
Tabela 17. Resultados previstos pelo modelo para uma Prefeitura Municipal.....	54

SUMÁRIO

1. INTRODUÇÃO	10
2. REFERENCIAL TEÓRICO.....	13
2.1 ÓRGÃOS , DESPESA PÚBLICA E O TRIBUNAL DE CONTAS	13
2.1.1 Estágios da execução da despesa.....	13
2.1.2 Classificação da despesa.....	14
2.1.3 Órgãos e Entidades Públicas Municipais.....	18
2.1.4 Tribunal de Contas do Estado De São Paulo	19
2.2 APLICAÇÃO DE MACHINE LEARNING.....	19
2.2.1 O que é Machine Learning (ML)?	19
2.2.2 Support Vector Machine (SVM)	20
2.2.3 LightGBM (LGBM)	20
2.2.4 Random Forest (RF)	20
2.2.5 Métricas para calcular a eficiências dos modelos.....	22
2.2.6 Importância das variáveis no modelo Random Forest.....	22
2.2.7 Por que utilizar aprendizado de máquina?.....	23
2.2.8 Escolha das ferramentas de análise de dados	24
2.2.9 Scikit-learn	25
2.3 AMOSTRAGEM EM AUDITORIA E ANÁLISE DO MODELO DO TCE-SP.....	26
2.3.1 Amostra de auditoria.....	26
2.3.2 Descrição do modelo utilizado pelo TCESP	27
3 METODOLOGIA.....	28
3.1 DESENHO DA PESQUISA.....	29
3.2 ESCOLHA DAS VARIÁVEIS	30
3.3 COLETA DE DADOS	31
3.4 PROCESSAMENTO DOS DADOS	33
4 RESULTADOS.....	35
4.1 VARIÁVEIS UTILIZADAS.....	35
4.1.1 Variáveis demográficas e estruturais	37
4.2 VARIÁVEIS ECONÔMICAS E FISCAIS.....	38
4.2.1 Variáveis educacionais	40
4.2.2 Variáveis relacionadas à saúde	41
4.3 ANÁLISES INICIAIS.....	41

4.4 TREINAMENTO INICIAL DOS MODELOS	44
4.5 AJUSTE DOS HIPERPARÂMETROS	46
4.6 TREINAMENTO DOS MODELOS COM OS HIPERPARÂMETROS OTIMIZADOS	48
4.7 ANÁLISE DA CORRELAÇÃO ENTRE AS VARIÁVEIS.....	50
4.8 TREINAMENTO DOS MODELOS EXCLUINDO AS VARIÁVEIS ALTAMENTE CORRELACIONADAS	51
5 PROPOSTA APLICADA	54
6 DISCUSSÃO	57
7 CONCLUSÃO.....	60
REFERÊNCIAS	62

1. INTRODUÇÃO

No contexto da Administração Pública, a auditoria desempenha um papel fundamental na promoção da transparência, da eficiência e da responsabilidade na aplicação dos recursos públicos (Tiozzi & Medeiros, 2024). O principal objetivo de uma auditoria independente, de acordo com a Norma Brasileira de Auditoria do Setor Público (NBASP) 200, é alcançar um nível razoável de segurança de que as demonstrações contábeis não possuem distorções relevantes causadas por erro ou fraude, para que o auditor possa emitir uma opinião sobre a adequação das mesmas à estrutura de relatório financeiro aplicável (Instituto Rui Barbosa, 2023). Para alcançar esse objetivo, o auditor deve executar procedimentos de auditoria apropriados, geralmente por amostragem, considerando a finalidade do procedimento e as características do conjunto de dados sobre o qual a amostra é selecionada (Conselho Federal de Contabilidade, 2009). A qualidade desses procedimentos pode ser melhorada pela utilização de inteligência artificial (IA), especificamente na identificação de áreas de risco e automação de tarefas repetitivas (Law & Shen, 2024).

O uso de ferramentas de análise de dados e IA em procedimentos de auditoria não é algo novo. Em 2004, o American Institute of Certified Public Accountants (AICPA) já havia colocado a mineração de dados como uma das dez melhores tecnologias do futuro, e o Institute of Internal Auditors (IIA) a listava entre as quatro prioridades de pesquisa (Chye Koh & Kee Low, 2004). Na literatura há diversos trabalhos demonstrando o uso de IA em auditoria, Ucoglu (2020), por exemplo, constatou que as maiores empresas de auditoria do mundo, as chamadas de *Big Four*, desenvolveram diversas ferramentas que são utilizadas para análise de dados, avaliação de riscos e extração de informações de documentos. Reforçando essa perspectiva, Mitan (2024) fez uma revisão bibliográfica sobre o impacto da IA em auditoria externa, concluindo que o uso de *machine learning* (ML) tem muitas possibilidades de aplicação para detecção de fraude, pois ela melhora a qualidade da auditoria aumentando a eficiência no processamento e análise dos dados.

Já no cenário brasileiro, Zaniboni e Montini (2019) realizaram uma revisão dos artigos em português com aplicações de ML em contabilidade e identificaram 31 artigos sobre auditoria, entre 1992 e 2018, demonstrando a tendência crescente de artigos por ano. Apesar do aumento anual no número de publicações, este mesmo estudo identificou que a literatura nacional privilegia a utilização de métodos como regressões, dados em painel e análise de

clusters. Técnicas mais recentes de aprendizado de máquina aplicadas à auditoria, como *Random Forest* (RF), *Support Vector Machine* (SVM) e *LightGBM*, ainda são pouco utilizadas (Zaniboni & Montini, 2019).

Acompanhando essa tendência, o Tribunal de Contas do Estado de São Paulo (TCE-SP) tem a disposição um algoritmo de ML com a técnica de RF para analisar os padrões de gastos dos órgãos municipais e prever os valores esperados para o exercício a ser auditado. O objetivo principal é gerar alertas para casos em que o valor executado pelos órgãos difere significativamente do valor previsto pelo modelo. No entanto, sua utilização ainda é facultativa, e a adesão por parte das equipes de auditoria tem sido baixa, indicando que aumentar a confiança na ferramenta é essencial para ampliar seu uso. Essa dificuldade não é restrita ao TCE-SP, Neves et al. (2019) realizaram entrevistas e aplicaram um questionário aos auditores do Tribunal de Contas da União (TCU), concluindo que, apesar de possuírem pelo menos quatro sistemas de IA disponíveis, o uso das ferramentas é limitado e os auditores ainda demonstram pouca confiança em relação às mesmas.

Diante desse cenário, o presente trabalho teve como objetivo aperfeiçoar o algoritmo de *machine learning* utilizado pelo TCE-SP e propor uma metodologia para aplicá-lo na seleção de uma amostra de auditoria, por meio do desenvolvimento de uma ferramenta intuitiva voltada ao uso pelos auditores. Já os objetivos específicos foram: (i) automatizar e melhorar o pré-processamento dos dados; (ii) identificar novas variáveis relevantes para o modelo e melhorar a qualidade das já utilizadas; (iii) identificar a técnica de ML mais adequada para os dados disponíveis; (iv) ajustar os hiperparâmetros do modelo para otimizar seu desempenho preditivo; (v) propor uma metodologia para utilização dos resultados do modelo na identificação de pontos de atenção, auxiliando na seleção de amostra pelas equipes de auditoria. Assim, a pergunta de pesquisa desse trabalho é: Como algoritmos de machine learning podem ser treinados e aplicados para identificação de distorções em despesas públicas, contribuindo para seleção de amostra nos órgãos de controle?

O desenvolvimento e utilização de ferramentas de ML pelo TCE-SP se justifica pela capacidade que elas possuem de potencializar a produtividade dos recursos humanos disponíveis (Mitan, 2024), permitindo que os auditores concentrem esforços em atividades com maior capacidade analítica. Essa necessidade torna-se ainda mais evidente diante do elevado número de órgãos públicos sob fiscalização, do expressivo volume de recursos financeiros envolvidos e da limitada quantidade de servidores disponíveis. De acordo com o Tribunal de Contas do Estado de São Paulo (2024b), o órgão contava com 1.860 servidores ativos em dezembro de 2024 para fiscalizar 644 municípios e o Estado de São Paulo. Ainda segundo o

Tribunal de Contas do Estado de São Paulo (2024a), o orçamento municipal consolidado dos 644 municípios fiscalizados em 2024 ultrapassou R\$ 208 bilhões. Além disso, o orçamento estadual fiscalizado superou os R\$ 328 bilhões no mesmo período, de acordo com a Secretaria da Fazenda do Estado de São Paulo (2024). Diante desse cenário, o uso de técnicas de ML surge como uma estratégia para ampliar a capacidade analítica do Tribunal, otimizando o uso dos recursos humanos e fortalecendo a fiscalização sobre um volume tão expressivo de recursos públicos.

O trabalho é dividido em sete capítulos, iniciando com a introdução que contextualiza o estudo e apresenta o problema de pesquisa, os objetivos e sua justificativa. Em seguida, o referencial teórico é dividido em três grandes seções que tratam dos órgãos públicos, da atuação do TCE-SP e das classificações de despesa; dos fundamentos e aplicações de ML; dos princípios de amostragem em auditoria além e descrição do modelo atualmente adotado pelo TCE-SP. O capítulo de metodologia detalha a estrutura da pesquisa, explicando as etapas de coleta, tratamento e análise dos dados. No capítulo de resultados são apresentados o treinamento dos modelos, os critérios de avaliação e as otimizações realizadas. A proposta aplicada recomenda uma estratégia para utilização do modelo como ferramenta de apoio ao planejamento de auditorias. Na discussão dos resultados faz-se uma análise de todo o processo realizado, comparando os resultados dos modelos e relacionando com trabalhos anteriores. Por fim, o trabalho se encerra com a apresentação das conclusões, limitações e sugestões para pesquisas futuras.

2. REFERENCIAL TEÓRICO

O referencial teórico deste trabalho está estruturado em três seções principais, cada uma abordando temas específicos relacionados à pesquisa desenvolvida.

A primeira seção, intitulada Despesa pública, Órgãos Públicos e a atuação do Tribunal de Contas, aborda as definições e classificações da despesa pública, os conceitos de órgãos e entidades públicas municipais e o papel do TCE-SP como órgão responsável pela fiscalização e controle da aplicação dos recursos públicos.

A segunda seção, Fundamentos e Aplicações de Machine Learning, apresenta os conceitos fundamentais de IA e ML e discute os aspectos relacionados à sua aplicação. Esta seção também discute as vantagens dos algoritmos de ML no contexto de auditoria pública, a importância da seleção, tratamento e interpretação das variáveis e do uso das ferramentas *Python*, *notebook jupyter* e da biblioteca *Scikit-learn*.

Por fim, a terceira seção, chamada Amostragem em Auditoria e Análise do Modelo Atual do TCE-SP, aborda a metodologia empregada na seleção de amostras em auditorias financeiras e os critérios de materialidade. Além disso, descreve o modelo preditivo atualmente utilizado pelo TCE-SP, destacando suas limitações.

2.1 ÓRGÃOS , DESPESA PÚBLICA E O TRIBUNAL DE CONTAS

2.1.1 Estágios da execução da despesa

De acordo com o Tesouro Nacional (2024), despesa orçamentária é definida como “o conjunto de dispêndios realizados pelos entes públicos para o funcionamento e manutenção dos serviços públicos prestados à sociedade” (p.97). A execução dessa despesa segue os três estágios previstos na Lei nº 4.320/1964: empenho, liquidação e pagamento (Brasil, 1964).

O art. 58 da mesma norma define que empenho é “o ato emanado de autoridade competente que cria para o Estado obrigação de pagamento pendente ou não de implemento de condição” (Brasil, 1964). É importante ressaltar que o artigo 60 veda a realização da despesa sem prévio empenho, visando garantir que todas as despesas do órgão passem por essa fase. (Florêncio, 2022) explica que o empenho constitui a reserva de uma parcela do orçamento público para atender dívida reconhecida. A formalização do empenho é realizada por meio da emissão da Nota de Empenho, documento essencial à execução orçamentária, no qual devem

constar, obrigatoriamente, a identificação do credor, a discriminação da despesa, o valor empenhado e demais informações pertinentes ao controle e à fiscalização da execução orçamentária e financeira (Gama, 2012).

A liquidação refere-se à verificação e apuração do direito adquirido pelo credor. Nesse estágio, é analisado se o fornecedor cumpriu integralmente suas obrigações contratuais, o que o torna apto a receber o pagamento, conforme estabelece o artigo 64 da Lei nº 4.320/1964 (Brasil, 1964).

O pagamento é o estágio final da execução da despesa orçamentária, definido no artigo 64 da lei 4.320/1964, e consiste na transferência efetiva de recursos financeiros do ente público contratante para o fornecedor, cumprindo as obrigações financeiras assumidas (Brasil, 1964).

2.1.2 Classificação da despesa

A Portaria Interministerial nº 163 (Brasil, 2001), estabelece padronização das classificações de receita e despesa com o objetivo de uniformizar procedimentos orçamentários entre União, Estados, Distrito Federal e Municípios, de acordo com o que está previsto na Lei de Responsabilidade Fiscal – LC nº 101 (Brasil, 2000). Essa classificação é obrigatória no âmbito da União, dos Estados, do Distrito Federal e dos Municípios e permite a consolidação nacional das despesas públicas (Gama, 2012). Essa obrigatoriedade também é explicitada no Manual de Contabilidade Aplicada ao Setor Público, publicado pelo Tesouro Nacional (2024): “O conjunto de informações que constitui a natureza de despesa orçamentária forma um código estruturado que agrega a categoria econômica, o grupo, a modalidade de aplicação e o elemento. Essa estrutura deve ser observada na execução orçamentária de todas as esferas de governo” (p.79).

Essa codificação da classificação da natureza da despesa é definida no artigo 5º da Portaria Interministerial nº 163 (Brasil, 2001), que descreve a estrutura completa da natureza da despesa como: c.g.mm.ee.dd, onde:

c = Categoria econômica;

g = Grupo de natureza da despesa;

mm = Modalidade de aplicação;

ee = Elemento de despesa;

dd = Subelemento (ou subitem), que é opcional.

A seguir, são apresentadas as definições e finalidades de cada item que compõe essa classificação, da categoria econômica até o subelemento, com base na legislação vigente e na literatura especializada.

O artigo 12, capítulo III, da Lei nº 4.320/1964 (Brasil, 1964) classifica as despesas de acordo com a categoria econômica em Correntes e de Capital, assim definidas por Florêncio (2022):

“3- Despesas Correntes: são aquelas destinadas à manutenção e funcionamento dos serviços públicos gerais anteriormente criados na administração. Classificam-se nesta categoria todas as despesas que não contribuem, diretamente, para a formação ou aquisição de um bem de capital.

4- Despesas de Capital: são as despesas que contribuem, diretamente, para a formação ou aquisição de um bem de capital (investimentos-material permanente e obras, inversões financeiras aquisição de imóveis, ou de bens de capital já em utilização). De maneira simples pode-se considerar que despesas correntes são aquelas destinadas ao custeio da máquina pública já existente, por exemplo pagamento de salários dos servidores, material de expediente, manutenção dos veículos, merenda para as escolas, entre outros. Já as despesas de capital podem ser consideradas como investimentos para expansão dos serviços disponibilizados para a população, por exemplo a construção de uma escola, de uma avenida, de um hospital, etc.” (p. 32).

O grupo de natureza da despesa agrupa as despesas em grupos com características semelhantes: Pessoal e Encargos Sociais, Juros e Encargos da Dívida, Outras Despesas Correntes, Investimentos, Inversões Financeiras e Amortização da Dívida (Tesouro Nacional, 2024). A modalidade de aplicação é uma informação gerencial que tem por finalidade indicar se os recursos são aplicados diretamente por órgãos ou entidades no âmbito da mesma esfera de Governo ou por outro ente da Federação e suas respectivas entidades (Tesouro Nacional, 2024). O elemento de despesa é definido pelo artigo 3º, §5º da Portaria Conjunta STN/SOF/ME: “O elemento de despesa tem por finalidade identificar os objetos de gasto, tais como vencimentos e vantagens fixas, juros, diárias, material de consumo, serviços de terceiros prestados sob qualquer forma, subvenções sociais, obras e instalações, equipamentos e material permanente, auxílios, amortização e outros de que a administração pública se serve para a consecução de seus fins” (Ministério da Economia, 2021).

O último nível da estrutura (dd - Subelemento) pode ser definido conforme as necessidades dos entes federativos, conforme estabelecido pela Portaria Conjunta STN/SOF/ME em seu artigo 3º, §5º: "É facultado o desdobramento suplementar dos elementos de despesa para atendimento das necessidades de escrituração contábil e controle da execução orçamentária". Ou seja, os entes públicos (União, Estados, Distrito Federal e Municípios) podem criar subelementos (também chamados de subitens) que serão classificados em cada um dos elementos de despesa para detalhar ainda mais os gastos, conforme suas necessidades e particularidades gerenciais, desde que esses desdobramentos respeitem a estrutura geral estabelecida pela portaria para que não comprometam a uniformização necessária para a consolidação das contas públicas.

O TCE-SP padroniza os subitens de despesa que devem ser utilizados pelos Municípios que fiscaliza nos balancetes contábeis enviados via sistema Audesp. A classificação de despesa até o nível de subitem é definida anualmente e fica disponível na página de documentação do sistema¹ classificada como tipo de documento: Plano de Contas no arquivo denominado Anexo II – Tabelas de Escrituração Contábil – Auxiliares (Tribunal de Contas do Estado de São Paulo, 2023a). De posse dessas tabelas de codificação, os próprios Órgãos Municipais e Entidades da administração indireta classificam seus gastos.

O modelo desenvolvido por este trabalho utiliza esse nível mais detalhado de desdobramento da classificação da despesa, por se tratar da estrutura que proporciona a maior granularidade das informações orçamentárias e permite uma identificação mais precisa do objeto da despesa pública. Para ilustrar essa classificação, pode-se citar o desdobramento do Elemento de Despesa 36 – Outros Serviços de Terceiros – Pessoa Física, o qual é subdividido em diversos subitens que especificam com mais detalhe o objeto dos serviços contratados.

Tabela 1. Primeiros 10 desdobramentos do elemento 36 utilizados no exercício de 2023

Codificação	Descrição do Código	Função
3.3.90.36.01	Condomínios	Registra o valor das despesas realizadas com taxas condominiais conta do locatário, quando previstas no contrato de locação.
3.3.90.36.02	Diárias a colaboradores eventuais no país	Registra o valor das despesas realizadas com diárias pagas a prestadores de serviços, de caráter eventual, sem vínculo com a administração pública (no país).

¹ <https://www.tce.sp.gov.br/audesp/documentacao>

Codificação	Descrição do Código	Função
3.3.90.36.03	Diárias a colaboradores eventuais no exterior	Registra o valor das despesas realizadas com diárias pagas a prestadores de serviços de caráter eventual, sem vínculo com a administração pública (no exterior).
3.3.90.36.04	Comissões e corretagens	Registra o valor das despesas realizadas com comissões e corretagens decorrentes de serviços prestados por: corretores, despachantes, leiloeiros e outros.
3.3.90.36.05	Direitos autorais	Registra o valor das despesas realizadas com direitos autorais sobre obras científicas, literárias ou em que a divulgação seja de interesse da entidade.
3.3.90.36.06	Serviços técnicos profissionais	Registra o valor das despesas realizadas com serviços prestados por profissionais técnicos, nas seguintes áreas: administração, advocacia, arquitetura, contabilidade, economia, engenharia, estatística, informática e outros.
3.3.90.36.07	Estagiários	Registra o valor das despesas realizadas com serviços prestados por estudantes na condição de estagiários ou monitores.
3.3.90.36.08	Bolsa de iniciação ao trabalho	Registra o valor das despesas realizadas com remuneração a candidatos participantes de curso de formação para o exercício de cargo decorrente de concurso público.
3.3.90.36.11	Pró-labore a consultores eventuais	Registra o valor das despesas realizadas com pró-labore a consultores eventuais.
3.3.90.36.13	Conferências, exposições e espetáculos	Registra o valor das despesas realizadas com o pagamento direto aos conferencistas e/ou expositores pelos serviços prestados.

Fonte: Tribunal de Contas do Estado de São Paulo (2023a)

Os dados de execução orçamentária, informados pelos órgãos públicos via balancetes contábeis, podem ser acessadas no portal eletrônico do Tribunal de Contas do Estado de São Paulo, que disponibiliza de forma pública e gratuita as informações prestadas desde 2008 (Tribunal de Contas do Estado de São Paulo, 2024a). É importante ressaltar o fato de que são informações declaradas pelos Municípios, conforme nota de rodapé do portal: “As informações constantes neste banco de dados foram fornecidas pelo próprio Município e serão objeto de

verificação pela fiscalização ordinária. Eventuais divergências serão conhecidas mediante a publicação do Parecer Prévio, correspondente à conta prestada”.

2.1.3 Órgãos e Entidades Públicas Municipais

Em cada município existem diversos órgãos e entidades que executam a despesa pública. A prefeitura municipal é o órgão central do poder executivo municipal responsável pela administração e gestão dos serviços públicos locais com autonomia administrativa, financeira e orçamentária para gerir os recursos municipais, conforme estabelece a Constituição Federal de 1988, em seus artigos 18 e 29 a 31 (Brasil, 1988).

As Entidades de Previdência são aquelas responsáveis pela gestão dos regimes próprios de previdência social (RPPS), voltados para os servidores públicos municipais. Esses regimes garantem aposentadorias e pensões aos servidores e são geridos por autarquias, fundações ou institutos específicos conforme estabelece a Lei nº 9.717/1998 (Brasil, 1998).

As Autarquias são entidades da administração pública indireta, criadas por lei, com personalidade jurídica própria e autonomia administrativa, financeira e patrimonial. Exercem atividades típicas do Estado, como fiscalização, regulação e prestação de serviços públicos, conforme definido no artigo 5º, inciso I, do Decreto-Lei nº 200 (Brasil, 1967).

As Câmaras Municipais são órgãos do poder legislativo municipal, composto por vereadores eleitos, responsáveis por elaborar leis locais e fiscalizar o poder executivo, de acordo com os artigos 29 e 31 da Constituição Federal (Brasil, 1988).

As Fundações Públicas são pessoas jurídicas de direito público ou privado, criadas pelo Estado para desempenhar atividades de interesse público, como educação, saúde e pesquisa, sem finalidade lucrativa. Sua criação e regime jurídico estão disciplinados pelo artigo 5º, inciso IV, do Decreto-Lei nº 200 (Brasil, 1967).

As Empresas Públicas são entidades da administração indireta, com personalidade jurídica de direito privado, criada para a prestação de serviços públicos ou exploração de atividades econômicas pelo Estado. Diferentemente das sociedades de economia mista, seu capital é 100% público, conforme previsto na Lei nº 13.303 (Brasil, 2016).

As Sociedades de Economia Mista são sociedades anônimas cujo controle acionário pertence majoritariamente ao poder público, mas também possuem participação de investidores privados. São criadas para explorar atividades econômicas ou prestar serviços públicos (Lei nº 13.303/2016), também nos termos da Lei nº 13.303 (Brasil, 2016).

2.1.4 Tribunal de Contas do Estado De São Paulo

O TCE-SP foi criado em 1921, e é um órgão auxiliar ao poder legislativo que tem como missão, de acordo com seu plano estratégico institucional, “Fiscalizar e orientar, por meio da atuação preventiva e corretiva e da avaliação de atos e resultados, para que os recursos públicos sejam utilizados de maneira adequada e transparente, em benefício da sociedade” (Tribunal de Contas do Estado de São Paulo, 2022, p. 10). Sua competência está definida no artigo 1º de seu regimento interno (Tribunal de Contas do Estado de São Paulo, 2023b):

“O Tribunal de Contas do Estado de São Paulo, órgão destinado à fiscalização contábil, financeira, orçamentária, operacional e patrimonial do Estado e de seus Municípios, auxiliar do Poder Legislativo no controle externo, tem sua sede na cidade de São Paulo e jurisdição em todo o território estadual.”

No tocante à fiscalização dos governos municipais, o artigo 33, inciso XIII, da Constituição do Estado de São Paulo estabelece que compete ao TCE-SP: “emitir parecer sobre a prestação anual de contas da administração financeira dos Municípios, exceto a dos que tiverem Tribunal próprio” (São Paulo, 1989). Isso significa que o TCE-SP tem jurisdição sobre 644 municípios do estado (excetuando-se apenas o município de São Paulo que possui Tribunal de Contas próprio).

2.2 APLICAÇÃO DE MACHINE LEARNING

2.2.1 O que é Machine Learning (ML)?

Em sua essência, ML é uma tecnologia que permite aos computadores identificar informações relevantes e padrões ocultos em grandes conjuntos de dados, encontrando tendências, prevendo resultados e tomando decisões sem a necessidade de serem programados explicitamente (Hardiansyah et al., 2024)

Rahman et al. (2020) explicam que ML é um campo de pesquisa em constante expansão, cujo objetivo central é o desenvolvimento de modelos preditivos capazes de alcançar altos níveis de acurácia com base nos conjuntos de dados analisados.

Aurélien Géron (2019), em seu livro sobre a utilização prática de aprendizado de máquina, explica que existem métodos supervisionados e não supervisionados, sendo que no primeiro os dados para treino já possuem as soluções desejadas e no segundo caso elas não

estão disponíveis. Entre os métodos de aprendizado supervisionado foram escolhidos três algoritmos para este estudo: *Support Vector Machine* (SVM), *LightGBM* (LGBM) e *Random Forest* (RF).

2.2.2 Support Vector Machine (SVM)

Rahman et al. (2020) apresenta o algoritmo SVM como um método de aprendizado de máquina supervisionado com algoritmos utilizados para tarefas de classificação e regressão. Além de realizar classificações lineares, o SVM também pode executar classificações não lineares de forma eficiente, minimizando o erro de classificação (Rahman et al., 2020). Tojeiro (2024) coloca que as SVMs apresentam como principal vantagem a sua eficácia na classificação de dados em espaços de alta dimensionalidade (o número de variáveis explicativas é elevado), porém possui uma limitação importante que é o seu custo computacional: “conforme o tamanho do conjunto de dados vai crescendo, o tempo necessário para fazer os cálculos aumenta rapidamente” (p. 63).

2.2.3 LightGBM (LGBM)

LGBM é uma biblioteca de ML projetada para ser mais rápida e eficiente, especialmente com grande volume de dados, utilizando algoritmos otimizados que reduzem o tempo de treinamento e o uso de memória, sem perder precisão (Zhao et al., 2024). Essa performance é alcançada adotando uma estratégia de identificar e priorizar a ramificação que proporciona maior ganho ao modelo (Yan et al., 2021).

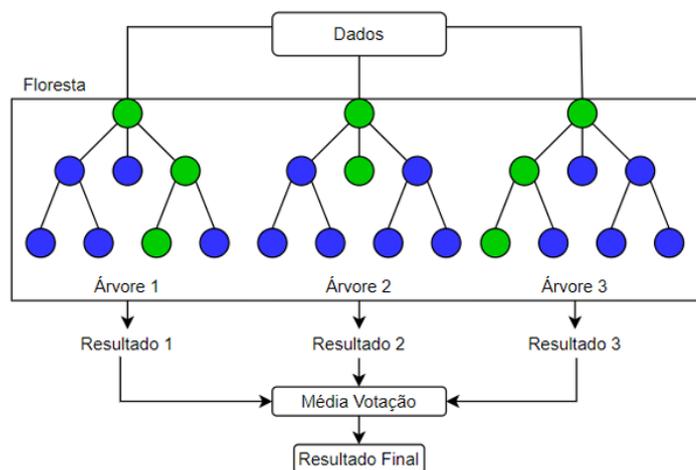
Zhang et al. (2023) e Ke et al. (2017), também apontam as vantagens associadas ao algoritmo *LightGBM*, principalmente devido à transformação das variáveis contínuas em intervalos discretos (*binning*), o que resulta em maior velocidade de treinamento e menor consumo de memória do computador. Donepudi et al. (2025) explicam que é um modelo que de boa acurácia, porque adota uma estratégia de divisão orientada por folhas (*leaf-wise*), permitindo a construção de árvores mais complexas e, conseqüentemente, modelos mais precisos. Li et al. (2024) também ressaltam a acurácia, confiabilidade e flexibilidade do *LightGBM*.

2.2.4 Random Forest (RF)

Breiman et al. (2017) definem RF como uma combinação de preditores baseados em árvores de decisão, em que cada árvore utiliza vetores aleatórios independentes e com a mesma distribuição. Segundo os autores, trata-se de uma ferramenta efetiva para previsão a partir de variáveis e, devido à lei dos grandes números, é uma técnica que não corre o risco de *overfit* (situação em que o modelo consegue resultados bons para os dados de treinamento, porém o desempenho nos dados de teste é baixo) (Breiman et al., 2017).

Tojeiro (2024) explica que o modelo de *Random Forest* “é usado tanto para classificação como também para regressão. Este classificador baseia-se na ideia de combinações de várias árvores de decisão em um único modelo preditivo” (p. 63). Em seu trabalho ele produz uma imagem para ilustrar esse processo de criação de várias árvores de decisão e a obtenção do resultado através da média da votação entre elas, conforme Figura 1.

Figura 1. Modelo de RF esquematizado



Fonte: (Tojeiro, 2024)

Trata-se de um modelo amplamente estudado na literatura, possui como principais vantagens sua capacidade de lidar com conjuntos de dados maiores, boa performance frente a *outliers* e redução do *overfitting*, porém apresenta como desvantagem sua alta complexidade que exige mais recursos e tempo computacional (Alnuaimi & Albaldawi, 2024). É um modelo capaz de avaliar novos dados com agilidade e obter uma boa capacidade preditiva (Zöller & Huber (2021). Sua alta capacidade preditiva e sua resistência ao *overfitting* também são características demonstradas em trabalhos anteriores, como os desenvolvidos por Montevechi et al. (2024) e Silva et al. (2025)

2.2.5 Métricas para calcular a eficiência dos modelos

O Erro Médio Absoluto (MAE - Mean Absolute Error) mede o erro médio entre os valores previstos e os valores realmente executados pelos órgãos, em unidades monetárias (Gigio, 2024; A. P. dos Santos et al., 2022). Willmott & Matsuura (2005) complementam que quanto menor o MAE, mais preciso é o modelo criado, tornando-o uma medida intuitiva e concluem recomendando que as comparações entre modelos sejam realizadas de acordo com essa métrica. Pode ser calculado pela fórmula:

$$MAE = \frac{1}{n} \sum_{i=1}^n |valor\ real - valor\ previsto|$$

O Erro Quadrático Médio (MSE - Mean Squared Error), Mishra (2021) explica que o MSE é uma das métricas mais utilizadas em tarefas de regressão por ser simplesmente a média das diferenças quadráticas entre o valor real e o valor previsto pelo modelo de regressão. Além disso, por elevar ao quadrado essas diferenças, o MSE penaliza até mesmo pequenos erros, o que pode levar a uma superestimação da inadequação do modelo (Mishra, 2021).

$$MSE = \frac{1}{n} \sum_{i=1}^n (valor\ real - valor\ previsto)^2$$

O Coeficiente de Determinação (R^2 - R-Squared) foi introduzido por Wright (1921) e quantifica o quanto a variável dependente é determinada pelas variáveis independentes (Chicco et al., 2021). O coeficiente de determinação tem como seu limite superior o valor 1 (que indicaria um ajuste perfeito do modelo) e não possui um limite inferior definido, porém o valor 0 ou negativo, indica que os dados não são explicados pelas variáveis independentes (Chicco et al., 2021).

2.2.6 Importância das variáveis no modelo Random Forest

As florestas aleatórias são modelos de ML preditivos que utilizam seleção aleatória de variáveis para construir um conjunto de árvores de decisão, porém a interpretação desses modelos é um desafio (Efron, 2020). A Medida de Importância da Variável (*Variable Importance Measure* - VIMP) é um método utilizado em modelos de RF para quantificar a relevância de cada variável independente no desempenho do modelo (Breiman et al., 2017).

Essa métrica ajuda a identificar quais variáveis exercem maior influência no poder preditivo do modelo, auxiliando na interpretação dos resultados e na seleção daquelas mais relevantes (de Oliveira, 2020).

A importância das variáveis pode ser calculada para modelos de RF utilizando o método da Permutação (*Permutation-Based VIMP*) (Teshale et al., 2025). Tradicionalmente, nesse método, o erro original do modelo é calculado utilizando um conjunto de dados original, em seguida a variável que está sendo analisada é embaralhada aleatoriamente em todas as observações, mantendo as outras constantes (Janitza et al., 2013). Variáveis importantes resultam em um aumento significativo do erro ou em uma diminuição do poder explicativo quando embaralhadas, indicando que o modelo depende delas para realizar suas previsões; por outro lado, variáveis que não são importantes apresentam pouca ou nenhuma alteração no erro ou no poder explicativo após a permutação, refletindo sua baixa relevância para o modelo (Teshale et al., 2025).

Wies et al. (2023) explicam que a importância da variável é a diferença entre o erro do modelo com os dados originais e o erro após a permutação da variável, se a variável não for associada ao resultado do modelo, a alteração da variável não terá influência no erro da previsão e, portanto, a diferença dos erros será próxima de zero. Apesar das vantagens desse método, a documentação oficial da biblioteca *Scikit-learn* alerta que esse método não calcula o poder preditivo específico da variável, apenas o quão importante ela é para o modelo que foi criado (Scikit-learn developers, 2023).

2.2.7 Por que utilizar aprendizado de máquina?

O uso da IA tem o potencial de aumentar a produtividade humana. O grande volume de dados que é gerado, armazenado e processado de maneira contínua e em alta velocidade está mudando o conceito e a natureza da auditoria (Brown-Liburd & Vasarhelyi, 2015). Os autores consideram que os sistemas de auditoria se tornarão um ecossistema completo, com processos automatizados e monitorados de forma contínua, melhorando a performance dos auditores humanos (Brown-Liburd & Vasarhelyi, 2015).

A adoção dessas ferramentas tem crescido de forma significativa em diversas áreas do conhecimento. No Brasil é possível encontrar na literatura diversas propostas de aplicação de aprendizado de máquina na administração pública. Dentre os estudos, destacam-se: a análise de prestação de contas de convênios (Carvalho, 2021); detecção de anomalias na execução de programas sociais, licitações e contratos (Nascimento, 2022); avaliação de risco fiscal (Savério,

2023); monitoramento do risco fiscal da dívida pública (Sá Filho, 2020); criação de um índice para priorização de municípios para auditoria com base em fatores de risco (Nascimento, 2022); e um modelo para classificação de risco para contratos públicos (Sales, 2017). De acordo com Borges et al. (2020), auditores internos brasileiros percebem a IA como uma aliada nas atividades de auditoria, que tende a substituir os métodos tradicionais, embora a maioria ainda possua apenas conhecimentos básicos sobre suas técnicas.

Outra questão relevante é se haveria outras ferramentas estatísticas mais adequadas para identificar observações discrepantes em gastos públicos. Stanišić et al. (2019) consideram que o desempenho comparativo dos classificadores depende fortemente da disponibilidade de informações previamente classificadas para treinar os modelos e explicam que, quando essas informações estão disponíveis, a regressão logística com efeitos mistos lida de forma eficaz com elas, fornecendo as previsões mais precisas. Na ausência de observações discrepantes já classificadas previamente, os algoritmos de aprendizado de máquina, particularmente os baseados em árvores de decisão, como RF e *LighGBM* produzem previsões mais precisas e utilizam melhor as informações disponíveis para realizar a previsão (Stanišić et al., 2019).

Apesar dos benefícios da utilização de técnicas de ML, seu uso também levanta preocupações adicionais. Njungle et al. (2025) alertam que existe risco de exposição de dados pessoais e confidenciais ao treinar modelos de ML, ressaltando a importância do desenvolvimento de ferramentas para preservação de privacidade. Costa (2021) propõe um grande debate sobre o uso de dados sensíveis em algoritmos de aprendizado de máquina, possíveis repercussões legais com base na Lei Geral de Proteção de Dados (LGPD) e a necessidade de consentimento ao usar essas informações. Rosa (2018) destaca que a mineração de dados com preservação da privacidade (*privacy preserving data mining* - PPDM), tornou-se um campo em evidência nos últimos anos, com o objetivo de proteger informações confidenciais contra divulgação não solicitada ou não autorizada, mas mantendo a utilidade dos dados. No entanto, como toda a análise deste trabalho é baseada em dados abertos e amplamente disponíveis ao público, o mesmo não enfrenta os riscos associados ao uso de informações sensíveis.

2.2.8 Escolha das ferramentas de análise de dados

A linguagem de programação *Python* foi escolhida por ser clara e simples, permitindo utilizar dados disponíveis de maneira ampla (Assis et al., 2021), com uma comunidade grande e ativa e ótimo suporte para bibliotecas (McKinney, 2018). Todo o código utilizado foi escrito

em *Notebooks Jupyter*, que funcionam como uma página interativa, acessada via navegador de internet. Nessa página é possível escrever o código necessário para as análises e também visualizar os resultados e gráficos (Pimentel et al., 2021). Uma vantagem ao utilizar *Notebooks Jupyter* é a racionalização do fluxo de trabalho com execução de pequenas partes do código de cada vez, facilitando a legibilidade e facilitando acesso aos resultados parciais (dos Santos, 2024). Em um trabalho envolvendo análise de dados em que haverá exploração dos dados, tentativas e erros e interações com diferentes fontes de dados, o uso da linguagem *Python* e de um *Notebook Jupyter* ajuda o usuário a executar suas tarefas mais rapidamente (McKinney, 2018). Diversos estudos já disponibilizaram seus *notebooks*, possibilitando que as análises sejam reproduzidas facilmente (Kluyver et al., 2016). Essa tendência crescente favorece a aplicação da metodologia proposta neste trabalho em outros conjuntos de dados (como informações contábeis de outros estados e até mesmo outros países) com pequenas adaptações ao código disponibilizado. Pimentel et al. (2021) ressaltam a utilidade dos *notebooks* para fazer ciência aberta: compartilhar o código e resultados para garantir a reprodutibilidade dos trabalhos.

2.2.9 Scikit-learn

O *Scikit-learn* é uma biblioteca da linguagem de programação *Python*, de código aberto, utilizada para desenvolver modelos de ML (Chakri et al., 2023). Em linguagens de programação, uma biblioteca é um conjunto de módulos e funções predefinidas que facilitam a análise de dados e permitem que o usuário execute tarefas complexas de forma simples, sem a necessidade de programar cada um dos passos necessários para alcançar o resultado (McKinney, 2018).

A biblioteca *Scikit-learn* fornece implementações avançadas de diversos algoritmos de ML amplamente conhecidos, mantendo uma interface intuitiva e atendendo à crescente demanda por análises estatísticas de dados por usuários que não são especialistas em programação, como biólogos e físicos (Pedregosa et al., 2011). Além de oferecer métodos eficientes de algoritmos de aprendizado de máquina, ela também disponibiliza ferramentas para pré-processamento de dados, seleção de modelos, validação e ajuste de hiperparâmetros (Pedregosa et al., 2011).

O ajuste de parâmetros é um processo essencial para a construção de modelos de aprendizado de máquina otimizados, garantindo que eles sejam capazes de gerar soluções eficazes para problemas de classificação ou regressão (Uludag, 2024). Para realizar essa

otimização, é necessário considerar todas as possíveis combinações dos parâmetros de configuração que influenciam seu desempenho (Hutter et al., 2020). A biblioteca *Scikit-Learn* possui o método *GridSearchCV* para realizar essa tarefa, testando todas as combinações possíveis dos parâmetros escolhidos pelo usuário para encontrar aquela que tem o melhor desempenho médio utilizando a validação cruzada (Scikit-learn developers, 2023). Trata-se de um processo de alternância entre as etapas de treinamento e validação ao longo de várias rodadas sucessivas com o objetivo de garantir que cada amostra do conjunto de dados tenha a oportunidade de ser utilizada para teste (Raschka, 2018). Essa técnica que divide os dados em subconjuntos (*folds*) e testa o modelo nessas diferentes divisões, melhorando a confiança em sua avaliação, pois evita que ele fique enviesado para um único conjunto de dados (Stock, 2025). Ademais, a validação cruzada é considerada a melhor técnica disponível para testar a capacidade de generalização de um modelo (Varoquaux, 2018).

2.3 AMOSTRAGEM EM AUDITORIA E ANÁLISE DO MODELO DO TCE-SP

2.3.1 Amostra de auditoria

De acordo com Francis (2024), embora as normas de auditoria exijam que os auditores coletem evidências suficientes para embasar a opinião de auditoria, não há uma regra clara que determine o ponto em que essa quantidade de evidências, necessária para sustentá-la, foi atingida. Como as normas de auditoria são, em grande parte, baseadas em princípios, elas oferecem ao auditor uma considerável flexibilidade na realização dessa avaliação (Francis, 2024).

Na escolha de uma amostra de auditoria, a materialidade é um componente essencial (Adhikari & Gårseth-Nesbakk, 2016). A NBASP 200 (Instituto Rui Barbosa, 2023) apresenta os conceitos de Materialidade Global (limite máximo aceitável para distorções nas demonstrações financeiras, definido com base em um percentual aplicado sobre um referencial apropriado) e Materialidade de Execução (um nível inferior à materialidade global, estabelecido para reduzir o risco de que, somadas as distorções, não detectadas e não corrigidas, ultrapassem a materialidade global) que devem ser aplicados à auditorias no setor público brasileiro. Diante desses conceitos, torna-se necessário definir um nível de materialidade aceitável para este trabalho.

A Norma Brasileira de Contabilidade Técnica de Auditoria – NBC TA 320 (Conselho Federal de Contabilidade, 2016) fundamenta como a materialidade deve ser definida: “4. A

determinação de materialidade pelo auditor é uma questão de julgamento profissional e é afetada pela percepção do auditor das necessidades de informações financeiras dos usuários” (p. 2). Leitão e Dantas (2016) realizaram uma pesquisa bastante abrangente com profissionais de Entidades de Fiscalização Superiores (EFS) internacionais, bem como profissionais do Tribunal de Contas da União (TCU) e da Controladoria Geral da União (CGU), buscando identificar quais seriam os parâmetros usados para definir a materialidade nas auditorias aplicadas ao setor público. Os resultados do estudo indicaram que, tanto nas EFS internacionais quanto entre os especialistas brasileiros, o parâmetro de materialidade mais adotado gira em torno de 2% do valor total analisado (Leitão & Dantas, 2016).

2.3.2 Descrição do modelo utilizado pelo TCESP

O modelo preditivo utilizado pelo TCE-SP considera um conjunto de variáveis que representam aspectos fiscais, econômicos, demográficos e educacionais dos municípios. No entanto, a coleta dos dados e seu tratamento eram bastante ineficientes. Uma de suas limitações é que a despesa dos anos anteriores era corrigida apenas pelo índice anual de inflação, desconsiderando o mês em que ocorreu a liquidação. Dessa maneira uma despesa realizada em janeiro e outra realizada em dezembro do mesmo ano recebiam a mesma correção pela inflação, ignorando a variação inflacionária ao longo do exercício. Ao aplicar o índice de inflação acumulada do exercício a uma despesa liquidada em dezembro, o valor resultante seria artificialmente maior do que o real. A consequência é que despesas realizadas em meses distintos recebiam a mesma correção inflacionária, o que distorcia a tendência histórica dos dados.

Além disso, outras variáveis eram bastante simplificadas. O PIB municipal era incluído no modelo como uma variável categórica, que registrava apenas qual a principal atividade econômica do município em questão (agropecuária, indústria ou serviços), sem utilizar os valores monetários. As quantidades de alunos e servidores públicos municipais eram tratadas como constantes para todos os anos. Essa simplificação visava agilizar a coleta de dados, em função da necessidade de velocidade na geração da análise preditiva logo no início do exercício seguinte à execução da despesa, porém distorcia os resultados do modelo, que não recebia os valores reais ano a ano para análise. Uma outra deficiência do modelo era a escolha das despesas que seriam consideradas para análise, pois optou-se por arbitrariamente selecionar aquelas com valores acima de R\$ 50.000,00 anuais e que ocorreram no mesmo município em pelo menos oito dos dez anos estudados.

3 METODOLOGIA

Este estudo caracteriza-se como uma pesquisa aplicada, com abordagem quantitativa, voltada à construção e validação de um modelo preditivo baseado em algoritmos de aprendizado de máquina. A proposta utiliza dados públicos reais para identificar despesas com comportamento atípico e demonstrar sua aplicação prática como ferramenta de apoio ao planejamento de auditorias, contribuindo para o direcionamento das análises por parte dos órgãos de controle.

A revisão da literatura revelou duas abordagens distintas para a utilização de variáveis no treinamento de modelos de *machine learning*, conforme discutido nos estudos de Stanišić et al. (2019) e Ying (2019).

Stanišić et al. (2019) realizaram uma extensa revisão de diversos modelos preditivos para geração de opinião de auditoria, desde aqueles que utilizavam estatística clássica (regressão probit e logística) até os que utilizam técnicas mais recentes de ML como árvores de decisão e redes neurais. Os autores argumentam que uma grande falha dos estudos está em não explorar integralmente os pontos fortes de cada técnica, ressaltando a necessidade de estudos que integrem melhor as vantagens das abordagens estatísticas e de aprendizado de máquina para aprimorar a predição de opiniões de auditoria. Em sua visão, os estudos que aplicaram técnicas de ML frequentemente se limitaram a um conjunto restrito de variáveis fundamentadas em teorias existentes, sem explorar o potencial dessas técnicas para lidar com um grande volume de preditores (Stanišić et al., 2019).

Em contraponto, Ying (2019) alerta que os modelos de ML tendem a considerar todas as variáveis disponíveis, mesmo que algumas delas tenham impacto insignificante no resultado. Em situações ainda mais problemáticas, certas variáveis podem atuar apenas como ruído, sem qualquer relevância para a predição. Esses casos podem ser solucionados ao selecionar apenas as variáveis mais relevantes, eliminando aquelas que não contribuem significativamente para o modelo.

A partir dos alertas de Stanišić et al. (2019) e Ying (2019), o treinamento dos modelos foi dividido em duas partes: (i) Treinamento de três modelos (RF, LightGBM e SVC) com todas as variáveis disponíveis, realizando a análise do custo computacional, erros e poder explicativo, a fim de identificar o modelo que melhor se adapta aos dados; (ii) Análise da correlação entre as variáveis, descartando aquelas de alta correlação para diminuir o risco de colinearidade e o

tempo de processamento dos modelos, avaliando o efeito da diminuição do número de variáveis nos erros e poder explicativo do modelo selecionado na etapa inicial.

3.1 DESENHO DA PESQUISA

Este estudo utiliza como variável dependente o valor liquidado anual por subelemento de despesa e por Prefeitura Municipal. Para prever esse valor, foram utilizadas como variáveis independentes um conjunto de informações sobre os municípios, descritas no capítulo 4, item 4.1 – Variáveis Utilizadas. Esses dados foram coletados de bases públicas oficiais e integrados em um único *dataset*, submetido a pré-processamento para normalização, tratamento de valores ausentes e codificação de variáveis categóricas. A base resultante foi então particionada em conjuntos de treinamento (dados de 2010 a 2022) e teste (dados referentes ao exercício em análise, 2023), garantindo a validação dos modelos em dados não conhecidos previamente pelos algoritmos.

A pesquisa foi conduzida em etapas sucessivas com o objetivo de desenvolver, avaliar e propor a aplicação prática de modelos preditivos para subsidiar o planejamento de auditorias públicas. Inicialmente, a base de dados foi consolidada a partir de diferentes fontes públicas, seguida pelo tratamento, correção monetária dos valores e aplicação de critérios de materialidade. Em seguida, realizou-se uma análise estatística descritiva com o apoio de gráficos de *boxplot*, permitindo compreender a distribuição das variáveis, identificando a dispersão, assimetria e orientando a normalização das variáveis.

Na etapa seguinte, foram treinados três algoritmos de aprendizado de máquina e seus desempenhos comparados com base em métricas de erro e capacidade explicativa. Os dois modelos com melhor desempenho foram otimizados por meio do método GridSearchCV, com validação cruzada, para identificar os melhores hiperparâmetros. Também foi conduzida uma análise de correlação entre as variáveis independentes, avaliando o efeito da exclusão daquelas altamente correlacionadas no tempo de treinamento e performance dos modelos.

O modelo que apresentou os melhores resultados foi utilizado para prever os valores de despesas por subelemento para cada uma das 644 prefeituras paulistas. As previsões foram comparadas aos valores efetivamente executados, e foram calculados intervalos de confiança empíricos de 95% para cada observação, utilizados para propor a aplicação prática desses resultados.

3.2 ESCOLHA DAS VARIÁVEIS

A previsão das despesas públicas municipais por meio de algoritmos de ML exige a seleção de variáveis que representem de forma abrangente os fatores que influenciam as despesas das Prefeituras Municipais. Os indicadores demográficos, econômicos e educacionais escolhidos neste estudo estão relacionados à arrecadação e à necessidade de alocação dos recursos públicos. A expectativa é que a combinação desses indicadores alimente o modelo com dados suficientes para gerar uma boa previsão do gasto público, baseado no consumo dos recursos disponíveis para atender a população.

A Tabela 2 apresenta as variáveis utilizadas no estudo acompanhadas de suas descrições.

Tabela 2. Variáveis utilizadas para treinamento dos modelos preditivos

Variável	Descrição
Cod_subelemento	Classificação da despesa ao nível de elemento de despesa ou subitem, dependendo de sua materialidade.
pop_total	População total do município, representando o número de habitantes residentes, conforme estimativas oficiais do IBGE.
vereadores	Número de vereadores que compõem a Câmara Municipal, definido conforme a população local pelo Art. 29 da Constituição Federal.
Area	Extensão territorial do município, expressa em quilômetros quadrados (km ²).
Agropecuária	Valor adicionado bruto do setor agropecuário no PIB municipal, incluindo atividades agrícolas, pecuárias, silvicultura, exploração florestal e pesca (valores em milhares de reais).
Indústria	Valor adicionado bruto do setor industrial no PIB municipal, abrangendo indústria extrativa, de transformação, construção civil e serviços industriais de utilidade pública (valores em milhares de reais).
Administração Pública	Valor adicionado bruto da administração pública no PIB municipal, incluindo serviços prestados por órgãos governamentais municipais, estaduais e federais (valores em milhares de reais).
Total (exclusive Administração Pública)	Valor total do PIB municipal sem considerar a administração pública, avaliando apenas a atividade econômica do setor privado (valores em milhares de reais).
Total geral	Valor total do PIB municipal, incluindo todos os setores da economia (agropecuária, indústria, serviços e administração pública) (valores em milhares de reais).
Impostos	Arrecadação de impostos sobre produtos, líquidos de subsídios, refletindo a carga tributária incidente sobre a produção e comercialização de bens e serviços (valores em milhares de reais).

Variável	Descrição
PIB (1)	Produto Interno Bruto. Soma de todos os bens e serviços finais produzidos no município em um período específico (valores em milhares de reais).
PIB per Capita (2)	PIB municipal dividido pela população total, refletindo a produção econômica média por habitante (valores em reais/habitante).
QT_MAT_BAS	Número total de matrículas na Educação Básica, abrangendo Educação Infantil, Ensino Fundamental e Ensino Médio - apenas rede municipal de ensino.
QT_MAT_INF	Número de matrículas na Educação Infantil, incluindo creche (0 a 3 anos) e pré-escola (4 e 5 anos) - apenas rede municipal de ensino.
QT_MAT_FUND	Total de matrículas no Ensino Fundamental, considerando anos iniciais (1º ao 5º ano) e anos finais (6º ao 9º ano) - apenas rede municipal de ensino.
QT_DOC_BAS	Número total de docentes atuantes na Educação Básica, incluindo Educação Infantil, Ensino Fundamental e Ensino Médio - apenas rede municipal de ensino.
QT_DOC_INF	Número de docentes que atuam exclusivamente na Educação Infantil - apenas rede municipal de ensino.
QT_DOC_FUND	Total de docentes no Ensino Fundamental, considerando anos iniciais e finais - apenas rede municipal de ensino.
receitas_correntes	Total das receitas correntes.
receitas_capital	Total das receitas de capital.
receita_total	Soma das receitas correntes e de capital.
leitos_total	Número total de leitos hospitalares disponíveis no município, incluindo os setores público e privado.
leitos_hab	Quantidade de leitos hospitalares por mil habitantes,
leitos_sus	Número de leitos hospitalares disponibilizados pelo Sistema Único de Saúde (SUS) no município. Oferta de atendimento público hospitalar.
leitos_sus_hab	Quantidade de leitos SUS por mil habitantes.
leitos_não_sus	Número de leitos hospitalares da rede privada ou não vinculados ao SUS.
leitos_naosus_hab	Quantidade de leitos não SUS por mil habitantes.

Fonte: Elaborado pelo autor com base nos dicionários de dados disponíveis nas bases consultadas (SEADE, INEP, IBGE).

3.3 COLETA DE DADOS

A principal base de dados deste estudo é composta pela execução orçamentária dos municípios do Estado de São Paulo, registrada por nota de empenho e disponível no Portal da Transparência Municipal (Tribunal de Contas do Estado de São Paulo, 2024a). A informação se origina nos balancetes contábeis informados pelos próprios órgãos municipais ao TCESP e,

conforme termo de isenção de responsabilidade disponível no sítio eletrônico, trata-se de informação declarada pelos próprios órgãos sem verificação por parte do Tribunal. A verificação se dá em um momento posterior e é divulgada com a publicação do parecer prévio. Essa avaliação posterior realizada pelo controle externo não foi considerada no estudo, pois são informações não estruturadas disponibilizadas no parecer prévio emitido em formato PDF e o tratamento necessário para sua utilização seria excessivamente custoso para esse estudo.

A análise de dados em documentos PDF representa uma barreira significativa para aplicação em modelos de ML, pois esse tipo de documento não possui uma estrutura padronizada. Converter esses arquivos PDF para um formato utilizável na modelagem exigiria técnicas avançadas de Processamento de Linguagem Natural (NLP) e Reconhecimento Óptico de Caracteres (OCR) (Costa, 2021).

Além dos dados orçamentários, foram realizadas consultas a bases de dados complementares para a obtenção de variáveis importantes para o modelo preditivo, incluindo:

- Fundação Sistema Estadual de Análise de Dados (SEADE): PIB Municipal, e leitos hospitalares disponíveis por exercício.
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP): Dados educacionais relacionados às matrículas e docentes na rede pública municipal.
- Instituto Brasileiro de Geografia e Estatística (IBGE): Utilizado para a obtenção da população total e área territorial dos municípios, além do índice de preços ao consumidor amplo (IPCA).

Os dados relativos ao IPCA mensal foram obtidos utilizando a API Sidra, mantida pelo IBGE, com os seguintes parâmetros de consulta, conforme Figura 2.

Figura 2. Parâmetros da consulta à API Sidra

```
# Parâmetros da consulta
table_code = "1737" #IPCA - Série histórica com número-índice,
#variação mensal e variações acumuladas em 3 meses,
# em 6 meses, no ano e em 12 meses (a partir de dezembro/1979)
territorial_level = "1" #Brasil(1)
ibge_territorial_code = "1" #Brasil(1)
variable = "63" #IPCA - Variação mensal (%) [janeiro 1980 a abril 2024] - casas decimais: padrão = 2, máximo = 2
period = "201001-202312" #Mês(533)
```

Fonte: Elaborado pelo autor

Nesse ponto do estudo é importante destacar a dificuldade na obtenção de dados em tempo hábil para o treinamento dos modelos preditivos. Especialmente quando se considera que a utilidade desse tipo de análise se dá principalmente no planejamento do trabalho de auditoria. No caso desse trabalho há uma curta janela para disponibilização desses dados, pois para que sua utilização ocorra de forma eficiente os auditores precisam que as previsões estejam disponíveis logo no início do exercício posterior ao da análise.

No entanto, as diferentes fontes de dados como IBGE, Fundação SEADE e INEP, possuem cronogramas distintos de atualização. Em alguns casos a defasagem temporal pode ser maior do que dois anos e impacta as informações disponíveis para utilização no modelo, impossibilitando o uso de variáveis potencialmente relevantes.

3.4 PROCESSAMENTO DOS DADOS

O primeiro passo para o processamento dos dados foi criar *scripts* na linguagem *Python* para acessar e consolidar os valores das despesas municipais, os valores do IPCA e demais dados utilizados neste trabalho. Trata-se de uma etapa bastante operacional, porém de extrema importância para garantir a celeridade na geração das análises em períodos futuros.

Com relação à despesa, considerando o objetivo central deste estudo (a identificação de padrões nos gastos públicos para prever despesas futuras), optou-se por utilizar os dados referentes à despesa liquidada. Trata-se da fase que realmente reflete uma obrigação reconhecida e consolidada, garantindo maior relação com o período a que efetivamente pertence a despesa, o que mitiga riscos de interpretações enviesadas ou inconsistentes. O uso dos valores empenhados ou pagos poderia comprometer o acompanhamento temporal dos gastos, dificultando a identificação de padrões consistentes, pois ambos podem estar distantes do efetivo recebimento da mercadoria ou prestação do serviço.

Também foi necessário criar o *script* para automatizar a correção monetária dos valores das despesas, que constam no banco de dados em seus valores nominais, para dezembro de 2023 utilizando o IPCA. A partir dos valores da inflação mensal de janeiro de 2010 a dezembro de 2023, foi calculada a inflação acumulada do mês da liquidação da despesa até dezembro de 2023, e esse valor foi aplicado a todas as despesas liquidadas mensais. Na Tabela 3 é apresentada uma amostra para ilustrar o resultado obtido.

Tabela 3. Amostra de valores corrigidos pelo IPCA até dezembro de 2023

Município	Órgão	Subelemento	ano_mes	vl_despesa	vl_corrigido
Jacupiranga	Câmara Municipal de Jacupiranga	31901302 - Contribuições previdenciárias - INSS	201007	5.057,32	11.011,74
Ubatuba	Prefeitura Municipal de Ubatuba	3903004 - Gás engarrafado	201111	18.612,00	37.415,82
Luiziânia	Câmara municipal de luiziânia	31901144 - Férias - abono pecuniário	201906	6.778,66	8.806,22
São Carlos	Fesc - Fundação Educacional São Carlos	33903022 - Material de limpeza e produtos de higienização	201311	3.312,60	5.966,84
Mirassol	Câmara Municipal de Mirassol	33903635 - Serviço de apoio administrativo, técnico e operacional	201709	240,00	334,96

Fonte: Elaborado pelo autor a partir do conjunto de dados de despesas municipais

Na Tabela 3 o `vl_despesa` é o valor contabilizado pelo órgão em determinado `ano_mes` (formato AAAMM) e o `vl_corrigido` representa esse valor atualizado pelo IPCA. A correção dos valores das despesas municipais pela inflação foi realizada para garantir a comparabilidade temporal dos dados e melhorar a precisão dos modelos preditivos de ML. Esse processo permite que as despesas sejam expressas em valores reais, eliminando distorções causadas pela perda do valor do dinheiro ao longo do tempo. A importância da correção monetária se dá porque, sem esse ajuste, o algoritmo poderia interpretar incorretamente que os gastos em 2010 foram significativamente menores do que o necessário para atender a mesma população em anos posteriores. Esse erro ocorreria porque os valores nominais das despesas daquele ano não refletem seu real poder de compra, levando o modelo a subestimar os gastos passados e, possivelmente, a gerar previsões distorcidas para períodos futuros. O mesmo processo foi utilizado para atualizar monetariamente os valores arrecadados mensais das receitas públicas. Como resultado, sempre que forem apresentados valores de despesas e receitas neste trabalho, trata-se de valores corrigidos pela IPCA até dezembro de 2023. Essa atualização monetária não foi necessária nos dados do PIB, pois a fundação SEADE já apresenta as informações em valores correntes.

Outro passo importante no pré-processamento dos dados foi a normalização dos valores numéricos, pois modelos como Máquinas de Vetores de Suporte (SVMs) são altamente sensíveis à escala das variáveis, o que pode influenciar sua capacidade de generalização e desempenho preditivo (de Amorim, Cavalcanti, & Cruz, 2023).

4 RESULTADOS

Este capítulo apresenta uma análise estatística descritiva do conjunto de dados utilizado, o treinamento dos modelos preditivos e a avaliação de seus desempenhos. Inicialmente foram calculadas as estatísticas descritivas e elaborados os gráficos de *boxplot* das variáveis, divididas em grupos temáticos, evidenciando a dispersão e assimetria presentes nas distribuições. Em seguida, o capítulo detalha a análise de materialidade e critérios para exclusão de subelementos pouco representativos. Com base nos dados tratados, três modelos de machine learning (*Random Forest*, *LightGBM* e *SVM*) foram treinados, avaliados e comparados, em suas versões padrão e com ajuste dos hiperparâmetros. Por fim, é discutido o impacto da exclusão de variáveis altamente correlacionadas no desempenho e na eficiência computacional dos modelos.

4.1 VARIÁVEIS UTILIZADAS

Com o objetivo de caracterizar o conjunto de dados utilizado neste estudo, a Tabela 4 apresenta as estatísticas descritivas das variáveis numéricas incluídas no modelo de ML referente ao exercício de 2023. Foram calculados a média, mediana, desvio padrão, valor mínimo e valor máximo para cada variável, de modo a proporcionar uma visão geral da distribuição e variabilidade dos dados.

Tabela 4. Estatísticas descritivas das variáveis utilizadas

Variável	Média	Mediana	Desvio Padrão	Mínimo	Máximo
pop_total	51.412,05	13.071,50	119.661,36	902,00	1.293.335,00
vereadores	10,75	9,00	3,24	9,00	34,00
area	382,56	280,50	317,03	3,00	1.978,00
Agropecuária	72.208,46	42.245,66	100.865,94	0,00	1.302.278,45
Indústria	666.953,25	66.388,05	2.011.432,07	1.440,39	27.735.676,16
Administração Pública	269.968,10	74.123,32	593.894,56	12.251,76	6.681.541,14
Total (exclusive Administração Pública)	1.567.424,23	194.330,69	5.112.670,78	11.599,36	68.268.461,36
Total geral	2.576.554,03	415.427,45	7.198.051,95	38.959,75	76.125.018,83
Impostos	455.627,82	35.631,78	1.633.884,44	-353.248,07	19.707.906,73

Variável	Média	Mediana	Desvio Padrão	Mínimo	Máximo
PIB (1)	3.032.181,85	448.682,99	8.775.319,45	40.362,73	95.832.925,57
PIB per Capita (2)	46.077,16	35.161,24	45.278,44	10.330,57	512.715,09
QT_MAT_BAS	4.917,46	1.564,00	9.974,08	80,00	97.183,00
QT_MAT_INF	1.883,16	553,00	3.896,16	32,00	34.995,00
QT_MAT_FUND	2.921,96	983,00	5.965,95	0,00	59.739,00
QT_DOC_BAS	345,89	118,00	635,94	7,00	5.627,00
QT_DOC_INF	172,84	52,00	343,41	4,00	3.132,00
QT_DOC_FUND	170,34	67,00	300,02	0,00	2.617,00
receitas_correntes	274.449.667,58	73.870.691,65	644.973.600,28	19.501.537,03	7.045.534.636,86
receitas_capital	12.071.387,55	3.721.122,56	31.130.204,70	0,00	480.767.407,18
receita_total	286.521.055,13	78.771.624,34	668.560.322,60	20.160.233,25	7.204.687.831,11
leitos_total	99,54	15,00	271,29	0,00	2.785,00
leitos_hab	1,37	0,65	2,25	0,00	27,92
leitos sus	60,32	10,00	154,38	0,00	1.425,00
leitos_sus_hab	0,93	0,44	1,70	0,00	21,72
leitos não sus	39,22	0,00	128,04	0,00	1.414,00
leitos_naosus_hab	0,44	0,00	1,01	0,00	9,63

Fonte: Elaborado pelo autor.

A análise estatística descritiva realizada sobre o conjunto de dados revela uma expressiva heterogeneidade entre os municípios do Estado de São Paulo. Observa-se que a variável população total apresenta uma média de 51.412 habitantes e uma mediana bastante inferior (13.071). Esses valores, juntamente com o elevado desvio padrão (119.661,36), evidenciam uma distribuição assimétrica e altamente concentrada em municípios de menor porte populacional. As variáveis relacionadas à saúde e educação apresentam uma distribuição bastante parecida com a da população e também uma grande diversidade em dimensão territorial, com as áreas variando de 3 km² a 1.978 km².

Na análise econômica, essa heterogeneidade também é encontrada. A indústria, apresenta o maior desvio padrão (R\$ 2.011.432,07) entre os setores econômicos analisados, demonstrando uma forte concentração da atividade industrial em poucos municípios. O mesmo se observa com o PIB, com média de R\$ 3.032.181,85, mediana de R\$ 448.682,99 e desvio padrão elevado (R\$ 8.775.319,45). Esses valores, juntamente com o PIB *per capita*, confirmam a concentração de riqueza em um número reduzido de municípios. Essas mesmas condições também se refletem nas receitas municipais, indicando uma desigualdade entre os municípios em termos de capacidade arrecadatória.

As variáveis foram organizadas em quatro grandes grupos: demográficas e estruturais, econômicas e fiscais, variáveis relacionadas à saúde e educacionais. Essa divisão temática permite uma compreensão mais clara das características estatísticas de cada conjunto, facilitando a identificação de padrões, assimetrias e possíveis *outliers*. Para cada grupo foi criado o gráfico de *boxplot*, esse tipo de visualização é especialmente útil para evidenciar a dispersão dos dados e a presença de valores extremos, informações relevantes tanto para o diagnóstico da base de dados quanto para orientar decisões de pré-processamento em modelos de ML (Leite, 2024).

4.1.1 Variáveis demográficas e estruturais

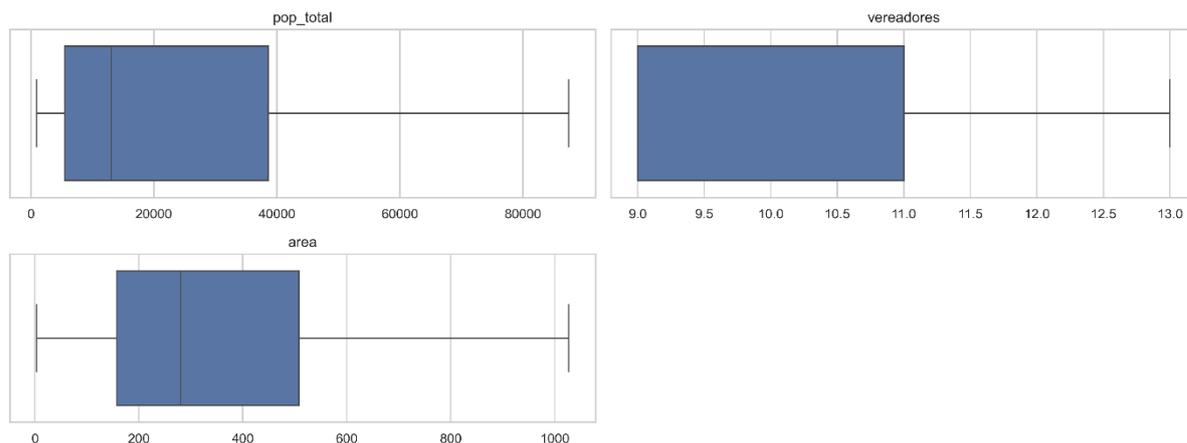
População total (*pop_total*) – O crescimento populacional, embora possa aumentar a eficiência do gasto (diminuindo o custo per capita com ganhos em escala) também causa um aumento nos gastos municipais devido à demanda por maior oferta de serviços públicos e investimentos (Lazaretti & Fochezatto, 2020).

Área territorial (*area*) – Municípios com maior extensão territorial podem ter custos operacionais mais elevados em infraestrutura e transporte (Silveira, 2022). Além disso, a relação entre a população e a área pode impactar a necessidade de investimentos em logística para distribuição de serviços.

Número de vereadores (*vereadores*) – O número de vereadores influencia o custo do poder legislativo municipal, pois Câmaras Municipais com mais parlamentares tendem a gerar despesas mais elevadas relacionadas a salários, assessores e infraestrutura administrativa (Silva et al., 2015).

A Figura 3 apresenta a distribuição das três variáveis demográficas e estruturais: população total, número de vereadores e área territorial dos municípios. Observa-se que a população apresenta uma assimetria positiva, indicando a presença de municípios com população significativamente maior que a mediana. A variável vereadores demonstra baixa variabilidade, com a maioria dos municípios concentrando-se entre 9 e 11 parlamentares. Por fim, a variável área apresenta também uma assimetria positiva, com poucos municípios apresentando extensões territoriais muito superiores à média e mediana.

Figura 3. Boxplot das variáveis demográficas e estruturais (Exercício de 2023)



Fonte: Elaborado pelo autor.

4.2 VARIÁVEIS ECONÔMICAS E FISCAIS

Variáveis relacionadas às receitas municipais: São três variáveis que representam os valores arrecadados por município, correspondendo às receitas correntes, de capital e a soma dessas duas. Essas variáveis representam os valores disponíveis para fornecer serviços e bens para a população (Tesouro Nacional, 2024).

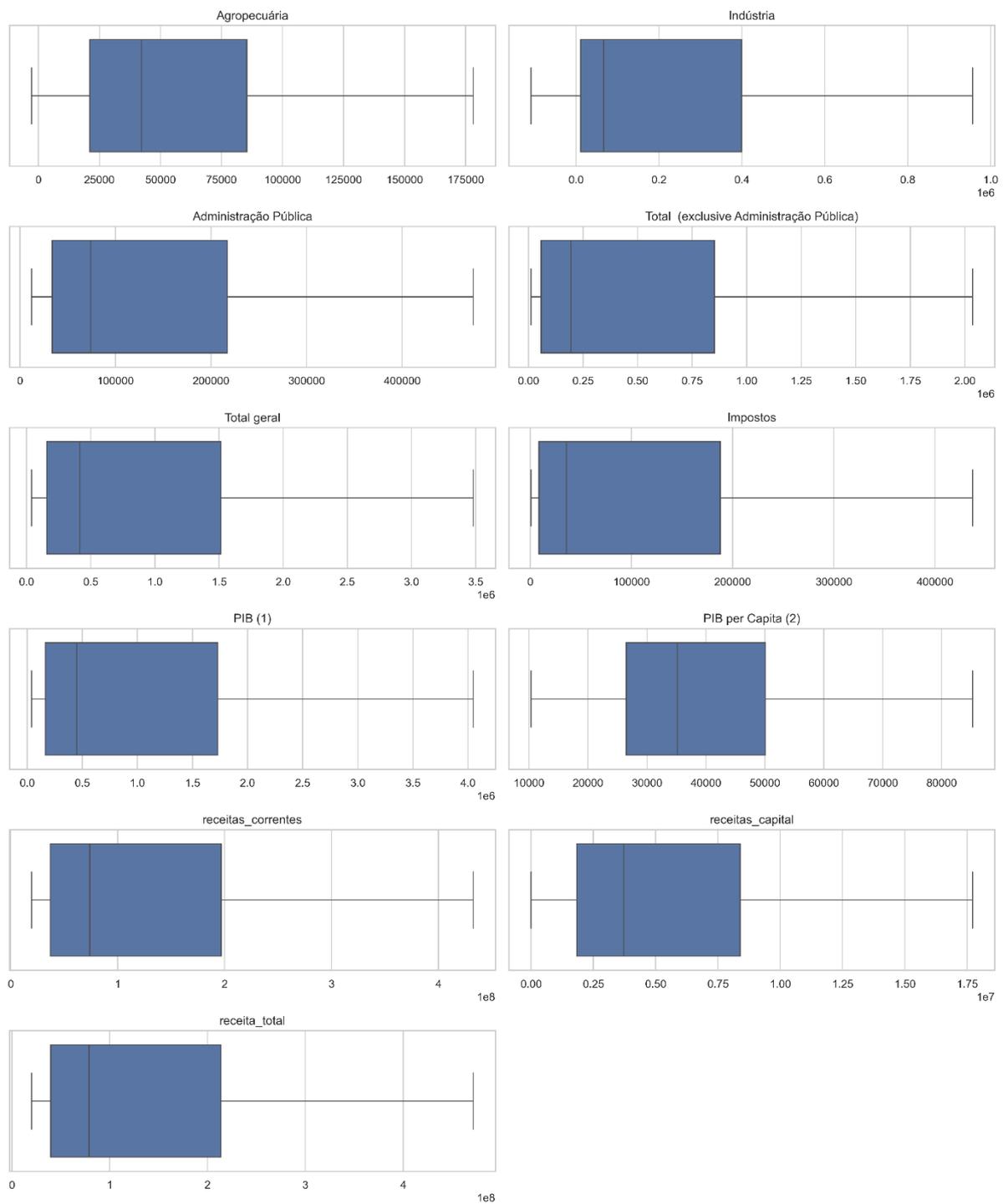
Produto Interno Bruto (PIB) e PIB per capita: O PIB municipal reflete a atividade econômica local e, conseqüentemente, o potencial de arrecadação tributária. Além disso, municípios mais desenvolvidos economicamente podem estar em estágios diferentes de priorização do gasto público (dos Santos et al., 2022).

Impostos arrecadados – A capacidade de arrecadação influencia diretamente o montante disponível para despesas, pois há uma forte relação entre os esforços fiscais e a estrutura de financiamento municipal (Moraes, 2006). Municípios que arrecadam mais impostos podem expandir investimentos e serviços públicos, enquanto aqueles com baixa arrecadação podem enfrentar restrições orçamentárias.

A Figura 4 demonstra que as variáveis econômicas e fiscais apresentam grande variabilidade e forte assimetria, refletindo a heterogeneidade entre os municípios estudados. Indicadores como Agropecuária, Indústria e Administração Pública revelam que a maioria dos municípios apresenta valores abaixo da mediana, enquanto poucos deles concentram níveis muito elevados de valor adicionado nesses setores. O PIB e o PIB *per Capita* reforçam essa desigualdade, com caudas longas à direita, indicando uma certa concentração da atividade econômica em poucos municípios. As variáveis *receitas_correntes*, *receitas_capital* e

receita_total seguem o mesmo padrão, demonstrando que a arrecadação municipal também é fortemente concentrada. Este padrão justifica a importância da normalização das variáveis em modelos preditivos.

Figura 4. Boxplot das variáveis econômicas e fiscais (Exercício de 2023)



Fonte: Elaborado pelo autor

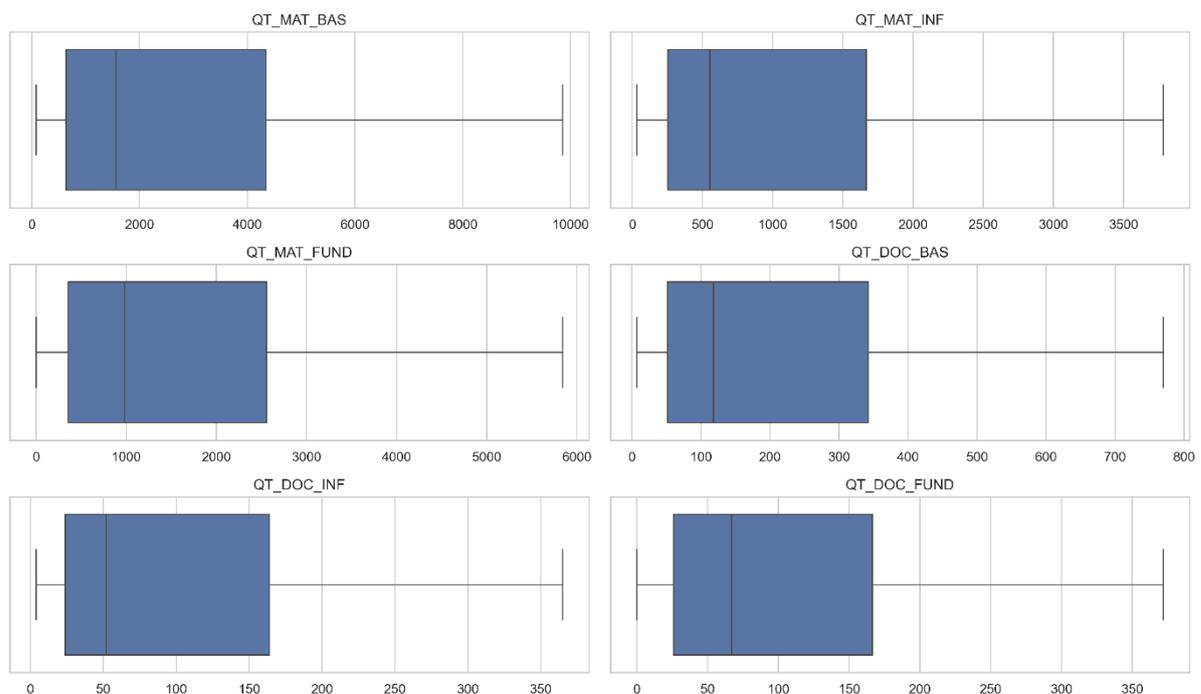
4.2.1 Variáveis educacionais

Matrículas na educação básica, infantil e ensino fundamental da rede municipal (QT_MAT_BAS, QT_MAT_INF, QT_MAT_FUND): A quantidade de alunos matriculados na rede municipal impacta diretamente os gastos com educação, incluindo infraestrutura escolar, alimentação e transporte. Espera-se que, com maior número de matrículas, as despesas relacionadas à educação também aumentem.

Número de docentes da rede municipal (QT_DOC_BAS, QT_DOC_INF, QT_DOC_FUND): A quantidade de professores em cada nível de ensino influencia os gastos com folha de pagamento, que geralmente representa um dos maiores gastos municipais da educação (Leite & Werneck, 2012).

A Figura 5 apresenta o conjunto de variáveis educacionais, abrangendo matrículas e número de docentes nos níveis de ensino básico, infantil e fundamental. Observa-se que todas as variáveis exibem distribuições assimétricas à direita, indicando a presença de municípios com valores significativamente superior à mediana. Esses resultados evidenciam que há uma grande disparidade na quantidade de alunos e docentes entre os municípios.

Figura 5. Boxplot das variáveis educacionais (Exercício de 2023)



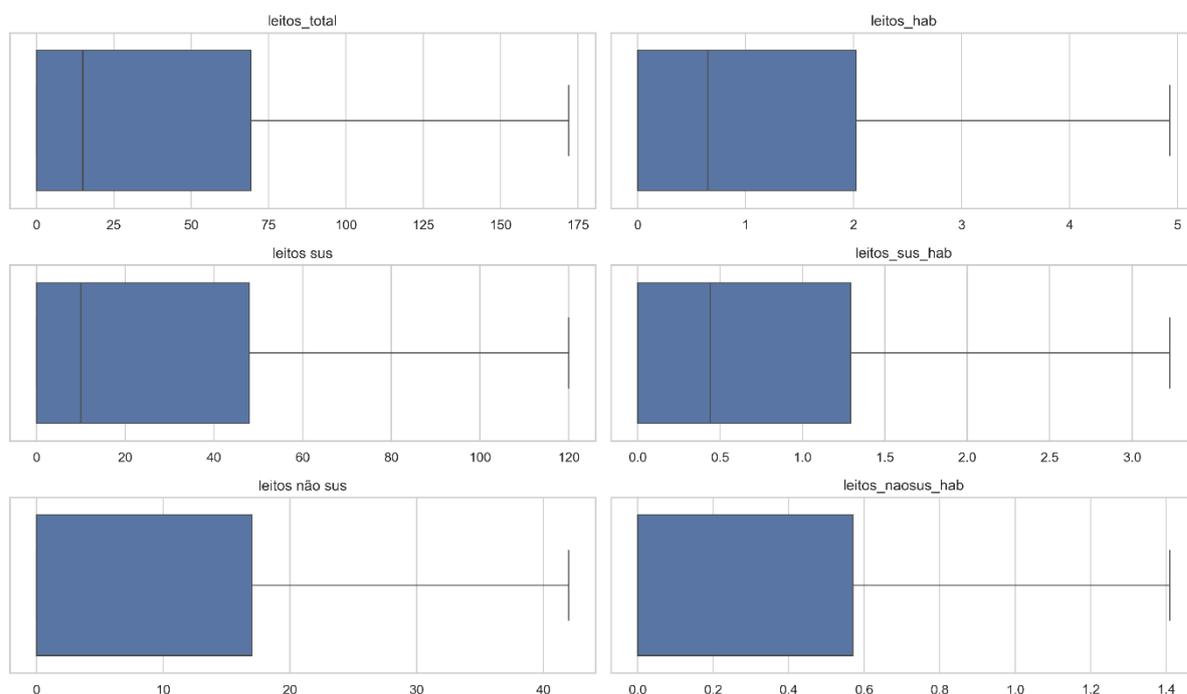
Fonte: Elaborado pelo autor

4.2.2 Variáveis relacionadas à saúde

Quantidade de leitos disponíveis (leitos_total, leitos_hab, leitos_sus, leitos_sus_hab, leitos_não_sus, leitos_naosus_hab): Nunes et al. (2024) abordaram a relação entre a capacidade de financiamento e os serviços de saúde ofertados pelos municípios, indicando que essas variáveis podem ser relevantes na construção de um modelo preditivo.

As variáveis relacionadas à saúde exibidas na Figura 6 apresentam uma distribuição variada. A variável leitos_total e seus desdobramentos (leitos_sus, leitos_não_sus) mostram assimetrias positivas acentuadas, com poucos municípios concentrando maior número absoluto de leitos hospitalares. Os indicadores por habitante (leitos_hab, leitos_sus_hab, leitos_naosus_hab) evidenciam melhor a disponibilidade relativa, com variância menor. Também são assimétricos, porém menos extremos.

Figura 6. Boxplot das variáveis da saúde (Exercício de 2023)

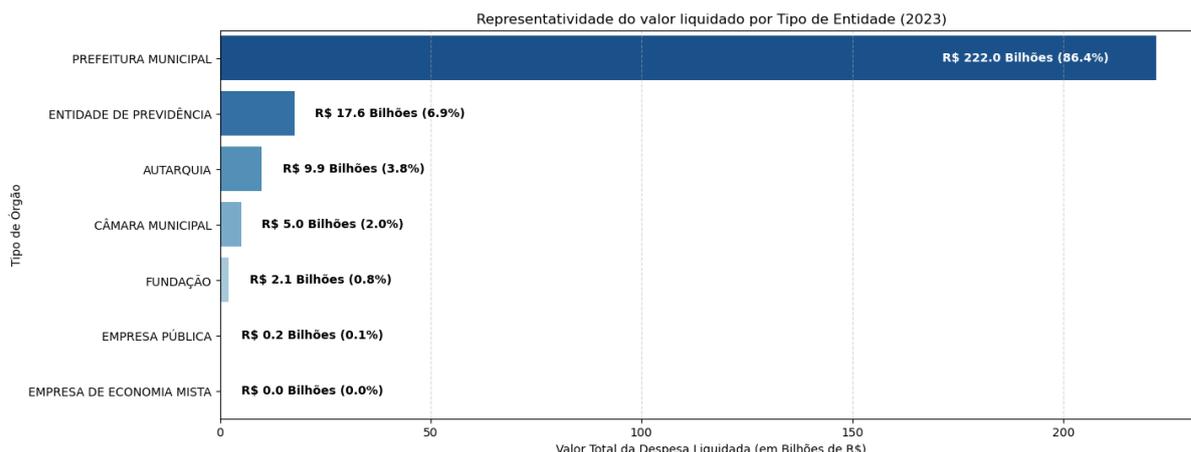


Fonte: Elaborado pelo autor

4.3 ANÁLISES INICIAIS

A primeira análise foi realizada para compreender a representatividade de cada tipo de órgão na composição da despesa total referente ao exercício de 2023.

Figura 7. Valor total da despesa liquidada em 2023 por tipo de entidade

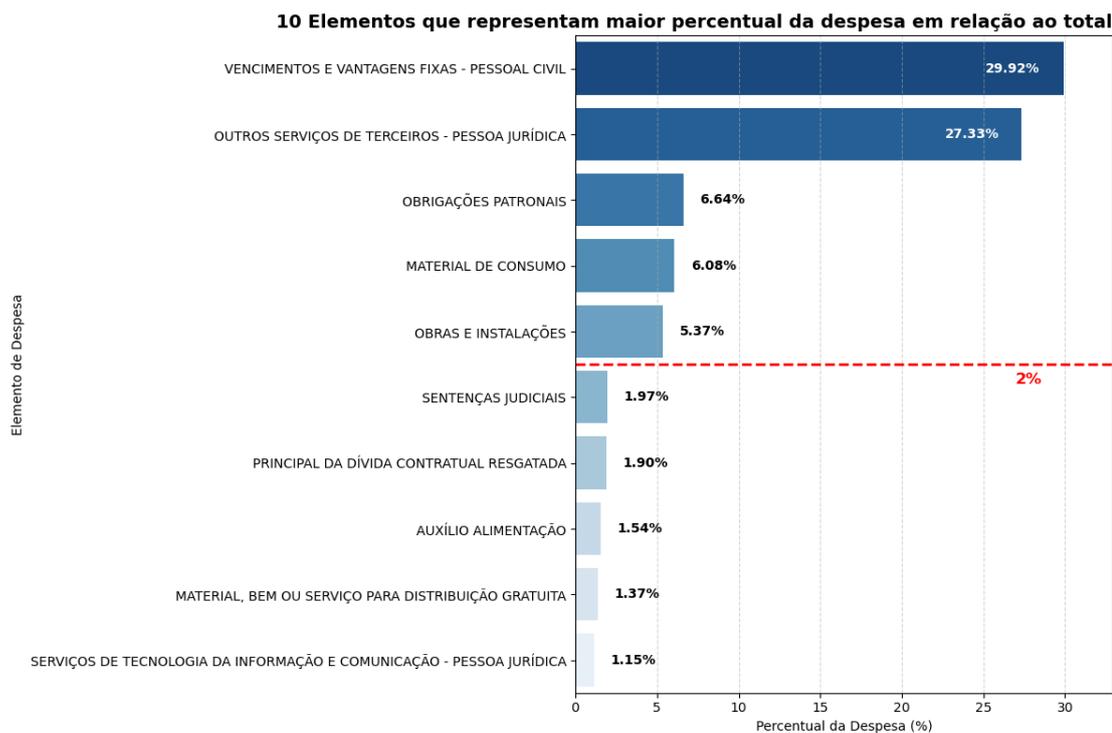


Fonte: Elaborado pelo autor

Observa-se pela Figura 7 que mais de 86% dos valores foram dispendidos pelas Prefeituras Municipais. Essa questão é relevante, pois dependendo do tipo de entidade analisada o padrão de gastos pode ser muito diferente. Uma Entidade de Previdência terá a maior parte de seus gastos em benefícios de aposentadorias e pensões, enquanto uma Autarquia ou Empresa pública terá seus gastos de acordo com seu âmbito de atuação e atribuições. Uma Autarquia que atua no ensino terá despesas em padrões diferentes do que uma Empresa de Transporte Urbano, por exemplo. Como o modelo de ML aprenderá com o padrão de gastos ao longo dos exercícios para prever a despesa no exercício em análise, misturar os dados de Entidades diferentes pode introduzir ruído, comprometendo sua precisão.

Tendo em vista que as Prefeituras Municipais concentraram mais de 86% do valor total das despesas liquidadas no exercício de interesse, os modelos apresentados neste capítulo foram desenvolvidos apenas para esse tipo de órgão público. Em seguida foi utilizado o critério de materialidade de execução de 2%, identificando as despesas que relevantes para a análise. É importante destacar que, adotando uma postura conservadora, foram desconsiderados apenas os subitens de despesa que, somados, representam menos de 2% da despesa total do exercício em análise. Partindo inicialmente da classificação por elemento de despesa, apenas 5 dos 41 elementos utilizados pelas Prefeituras Municipais em 2023 representavam mais do que 2% da despesa total.

Figura 8. Dez elementos que representam o maior percentual da despesa executada em 2023



Fonte: Elaborado pelo autor

A partir dos elementos menos representativos no critério de materialidade, foram identificadas aquelas despesas que poderiam ser agrupados ao nível de elemento, por exemplo o elemento 04 – Contratação por tempo determinado, que apresenta 10 subitens relacionados, conforme Tabela 5.

Tabela 5. Subdivisões do elemento de Despesa 04 – Contratações por tempo determinado

Código do Elemento / Subelemento	Descrição do Subitem	Valor Despesa	Percentual em relação ao total
0411	Salários - contrato temporário	358.855.160,83	0,20
0499	Outras contratações por tempo determinado	299.888.098,62	0,17
0418	Professores substitutos / visitantes	103.327.940,53	0,06
0416	Obrigações patronais - INSS - contrato temporário	49.845.058,83	0,03
0413	13º salário - contrato temporário	29.588.845,37	0,02
0415	Obrigações patronais	17.125.795,93	0,01

Código do Elemento / Subelemento	Descrição do Subitem	Valor Despesa	Percentual em relação ao total
0414	Férias - abono constitucional - contrato temporário	15.214.444,39	0,01
0417	Obrigações patronais - FGTS - contrato temporário	8.622.840,88	0,00
0400	Contratação por tempo determinado	4.559.548,26	0,00
0451	Adicionais de contrato temporário	4.293.951,30	0,00

Fonte: Elaborado pelo autor

Observa-se que os subitens deste elemento são muito pouco representativos e em geral são múltiplos de uma mesma base (13º salário, férias, obrigações patronais, entre outros, são todos múltiplos da remuneração básica do contratado). Por esse motivo optou-se por manter essa despesa agregada ao nível de elemento. O mesmo critério foi utilizado para os demais elementos, que representavam menos de 2% da despesa total para o exercício de 2023. O resultado desse agrupamento dos elementos menos representativos foi uma tabela com 219 subitens. Destes, o valor somado dos 118 menores representava menos de 2% da despesa liquidada total do exercício e eles foram excluídos da análise.

4.4 TREINAMENTO INICIAL DOS MODELOS

Inicialmente foram treinados três modelos, *Random Forest*, *LightGBM* e *SVM*, utilizando os hiperparâmetros padrão da biblioteca *SKLearn*. Os três modelos testados apresentaram tempos de treinamento distintos, conforme registrado na execução do método *fit()*, demonstrado na Figura 9.

Figura 9. Tempo de treinamento dos três modelos

```

1 model_rf.fit(X_train_scaled, y_train)
✓ 17m 31.6s Python
RandomForestRegressor ⓘ ?
RandomForestRegressor(random_state=42)

1 model_lgbm.fit(X_train_scaled, y_train)
✓ 1.5s Python
LGBMRegressor ⓘ
LGBMRegressor(random_state=42)

1 model_svm.fit(X_train_scaled, y_train)
✓ 183m 57.2s Python
SVR ⓘ ?
SVR ()

```

Fonte: Elaborado pelo autor

O modelo RF apresentou um tempo de treinamento intermediário, levando 17 minutos e 31 segundos para se ajustar aos dados de treinamento, enquanto o *LightGBM* demonstrou ser mais eficiente em termos computacionais, levando apenas 1,5 segundo para ser treinado. Já o SVM apresentou um tempo de treinamento extremamente elevado, levando 183 minutos e 57 segundos (mais de 3 horas). Esse tempo elevado para treinamento já era esperado, pois diferente dos modelos baseados em árvores de decisão, que segmentam os dados de forma eficiente, o treinamento de SVM pode ser computacionalmente caro (Santos, 2023), não sendo recomendado para conjuntos de dados muito grandes (Cortes & Vapnik (1995) *apud* Boateng et al. (2020)).

A avaliação dos erros dos modelos treinados foi realizada através do cálculo das três métricas MAE, MSE e R^2 , conforme Tabela 6.

Tabela 6. Avaliação dos modelos

Modelo	MAE (Valores em R\$)	MSE	R^2
Random Forest	1.442.655,41	$87,5 \times 10^{12}$	0,85

LightGBM	2.240.732,96	$114,9 \times 10^{12}$	0,81
SVM	3.575.407,74	$611,5 \times 10^{12}$	-0,02

Fonte: Elaborado pelo autor.

Avaliando os resultados iniciais dos modelos, observa-se que o *Random Forest* apresentou melhor desempenho em geral, com menor MAE (R\$ 1.442.655,41) e maior R^2 (0,85). Ou seja, possui maior poder explicativo e menor erro médio absoluto. Além disso, o MSE mais baixo indica que os erros cometidos pelo modelo foram menores em magnitude, o que demonstra uma boa capacidade de aprendizado. Com esse resultado inferiu-se que o modelo utilizando *Random Forest* conseguiu capturar mais padrões relevantes nos dados, fornecendo previsões mais próximas da realidade. Nas próximas seções foram realizados testes adicionais, como análises de importância de variáveis e avaliação de multicolinearidade, para entender melhor o impacto individual cada uma delas.

O modelo *LightGBM* teve um desempenho razoável, porém inferior ao da RF, com um MAE maior (R\$ 2.240.732,96) e um R^2 menor (0,81). Apesar de ainda apresentar um bom poder explicativo, ficou abaixo do modelo RF em todas as métricas de avaliação. Considerando a capacidade preditiva apresentada, continua sendo uma alternativa viável principalmente se a capacidade computacional disponível for baixa.

Por fim, o SVM apresentou desempenho inferior, com MAE de R\$ 2.240.732,96 e um MSE extremamente alto. O R^2 negativo (-0,02) indica que o modelo não conseguiu aprender com os padrões apresentados pelos dados, tendo um desempenho pior do que simplesmente utilizar os valores médios como previsão. Esse resultado em conjunto com o tempo excessivo necessário para seu treinamento foram os motivos pelos quais o SVM não foi considerado um modelo efetivo para a base de dados apresentada e foi descartado nos próximos passos.

4.5 AJUSTE DOS HIPERPARÂMETROS

Com a seleção dos dois melhores modelos (RF e *LightGBM*), o próximo passo foi encontrar os melhores hiperparâmetros utilizando GridSearchCV com validação cruzada da biblioteca *ScikitLearn*. Essa validação foi realizada utilizando o parâmetro $CV=3$ (*cross validation=3*), isso significa que o conjunto de dados foi dividido em três partes (*folds*), e o modelo foi treinado e avaliado três vezes, alternando os conjuntos de treino e teste para a ocorrência de *overfitting*. Trata-se de um processo importante, pois se o modelo fosse treinado em um único conjunto de dados fixo, poderia se escolher hiperparâmetros que só funcionam

bem naquele conjunto específico. Esse processo busca combinações ideais de parâmetros, testando diferentes configurações para identificar aquela que minimiza o erro da previsão. Os parâmetros aplicados no método para cada modelo foram:

Random Forest

- `n_estimators`: Número de árvores no conjunto (100, 300, 500).
- `max_depth`: Profundidade máxima das árvores (10, 20, None).
- `min_samples_split`: Mínimo de amostras para dividir um nó (2, 5, 10).
- `min_samples_leaf`: Mínimo de amostras em uma folha (1, 2, 4).

LightGBM

- `num_leaves`: Número de folhas na árvore (31, 50, 100).
- `learning_rate`: Taxa de aprendizado (0.01, 0.05, 0.1).
- `n_estimators`: Número de árvores (100, 300, 500).
- `max_depth`: Profundidade máxima (-1 significa sem limite).

O MAE foi escolhido como critério para definição dos melhores parâmetros por ser de fácil compreensão.

Após a execução do GridSearch, esses foram os melhores parâmetros encontrados para cada modelo, conforme Tabelas 7 e 8.

Tabela 7. Melhores parâmetros encontrados para o modelo Random Forest

Hiperparâmetro	Valor	Descrição
<code>max_depth</code>	None	Sem limite de profundidade, permitindo que as árvores cresçam até capturar todos os padrões possíveis.
<code>min_samples_leaf</code>	1	O menor número permitido de amostras em uma folha. Permite capturar mais variações nos dados.
<code>min_samples_split</code>	2	Os nós são divididos sempre que houver pelo menos 2 amostras, aumentando a flexibilidade do modelo.
<code>n_estimators</code>	100	Usa 100 árvores na floresta, equilibrando desempenho e tempo de treinamento.

Fonte: Elaborado pelo autor

Tabela 8. Melhores parâmetros encontrados para o modelo LightGBM

Hiperparâmetro	Valor	Descrição
<code>learning_rate</code>	0.1	Taxa de aprendizado equilibrada, permitindo ajustes moderados a cada iteração.
<code>max_depth</code>	-1	Sem limite de profundidade, deixando o modelo definir automaticamente o melhor tamanho das árvores.

Hiperparâmetro	Valor	Descrição
n_estimators	500	Número maior de árvores, garantindo aprendizado mais profundo e maior precisão.
num_leaves	100	Aumenta a complexidade das árvores, permitindo capturar padrões mais detalhados.

Fonte: Elaborado pelo autor

4.6 TREINAMENTO DOS MODELOS COM OS HIPERPARÂMETROS OTIMIZADOS

Os resultados da avaliação dos erros após o treinamento dos modelos com os parâmetros otimizados são apresentados nas Tabelas 9 e 10.

Tabela 9. Comparação dos erros dos modelos após otimização (*Random Forest*)

Métrica	Random Forest (Padrão)	Random Forest (parâmetros otimizados)
MAE	1.442.655,41	491.934,75
MSE	$87,59 \times 10^{12}$	$10,95 \times 10^{12}$
R ²	0,85	0,98

Fonte: Elaborado pelo autor

O modelo RF apresentou uma expressiva redução nos erros e um aumento considerável na capacidade explicativa. O valor do MAE diminuiu de R\$ 1.442.655 para R\$ 491.934, representando uma redução de 66%. Da mesma forma, MSE foi de $87,59 \times 10^{12}$ para $10,95 \times 10^{12}$, uma diminuição de 87,5%. O coeficiente de determinação (R²) aumentou de 0,89 para 0,98, indicando que o modelo agora é capaz de explicar 98% da variabilidade dos dados.

O aprimoramento das variáveis utilizadas e otimização dos parâmetros resultaram em uma melhora significativa no desempenho preditivo do modelo RF em comparação com a versão utilizada anteriormente pelo TCE-SP. A análise das métricas de erro sugere que a incorporação dessas informações adicionou valor ao modelo, tornando-o mais preciso na previsão da despesa. Comparando com as métricas do modelo anterior observamos uma

melhoria de 74,4% no MAE, 89,1% no MSE e 12,6% no R^2 (as métricas do modelo anterior foram 1.923.581,91; $100,51 \times 10^{12}$ e 0,87 respectivamente).

Tabela 10. Comparação dos erros dos modelos após otimização (*LightGBM*)

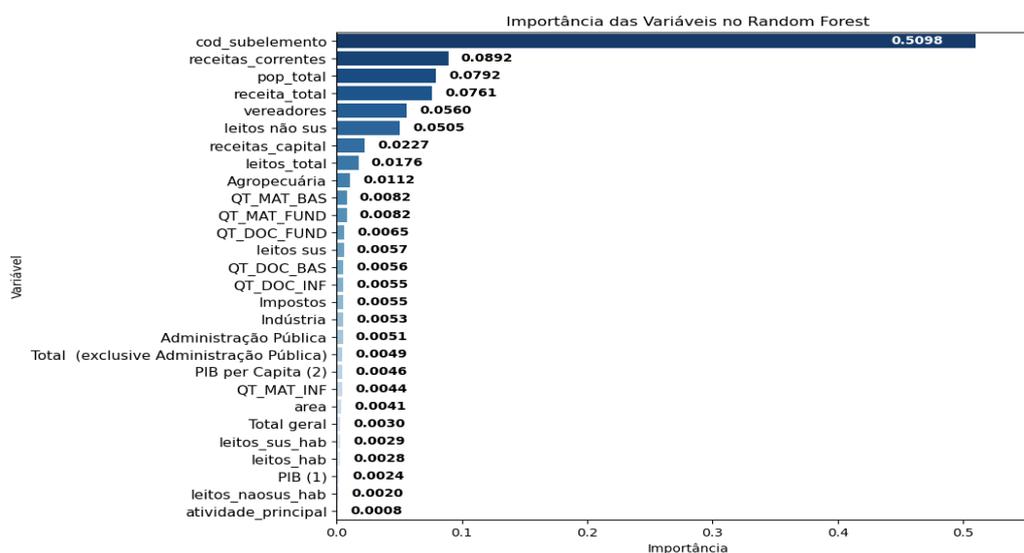
Métrica	LightGBM (Padrão)	LightGBM (parâmetros otimizados)
MAE	2.240.732,96	1.298.927,03
MSE	$114,99 \times 10^{12}$	$21,37 \times 10^{12}$
R^2	0,81	0,96

Fonte: Elaborado pelo autor

O modelo *LightGBM* também apresentou melhorias significativas, embora tenha permanecido com um desempenho inferior ao RF. O MAE reduziu de R\$ 2.240.732 para R\$ 1.298.927, representando uma queda de 42%. O MSE passou de $114,99 \times 10^{12}$ para $21,37 \times 10^{12}$, uma redução de 81,4%, enquanto o R^2 aumentou de 0,81 para 0,96, indicando que o modelo agora explica 96% da variabilidade da despesa. Apesar dessas melhorias, o erro médio ainda se manteve significativamente maior em comparação ao RF, sugerindo que este último modelo é mais indicado para previsão das despesas municipais.

Na etapa seguinte, realizou-se a análise de importância das variáveis, conforme apresentado na Figura 10.

Figura 10. Importância das variáveis no modelo RF otimizado



Fonte: Elaborado pelo autor

É possível observar um bom número de variáveis com importância menor do que 1%, indicando que sua influência nos erros do modelo é muito pequena.

4.7 ANÁLISE DA CORRELAÇÃO ENTRE AS VARIÁVEIS

A análise de correlação entre as variáveis educacionais é apresentada na Tabela 11.

Tabela 11. Correlação entre as variáveis que representam a quantidade de matrículas

Variável 1	Variável 2	Correlação
QT_MAT_BAS	QT_MAT_FUND	0,984
QT_MAT_BAS	QT_MAT_INF	0,959
QT_MAT_INF	QT_MAT_FUND	0,899

Fonte: Elaborado pelo autor

Os coeficientes de correlação entre a quantidade de matrículas na educação básica com as matrículas no ensino infantil e fundamental foi de 0,984 e 0,959, indicando uma relação linear forte, sugerindo que essas duas variáveis fornecem praticamente a mesma informação ao modelo. Ou seja, indica que a quantidade de matrículas na educação básica pode ser excluída do modelo sem que haja uma perda significativa na capacidade preditiva.

Ao analisar a quantidade de docentes, foi encontrada uma situação semelhante, o que possibilitou a exclusão da quantidade de docentes na educação básica, conforme Tabela 12.

Tabela 12. Correlação entre as variáveis que representam a quantidade de docentes

Variável 1	Variável 2	Correlação
QT_DOC_BAS	QT_DOC_FUND	0,96
QT_DOC_BAS	QT_DOC_INF	0,95
QT_DOC_INF	QT_DOC_FUND	0,82

Fonte: Elaborado pelo autor

A análise da correlação entre as variáveis relacionadas à quantidade de leitos hospitalares também revelou uma forte relação linear positiva. A variável leitos_total apresenta uma correlação de 0,960 com leitos_sus e 0,925 com leitos_não_sus, indicando possível redundância de informação no modelo preditivo.

A alta correlação entre leitos_hab e leitos_sus_hab levou a primeira a ser desconsiderada. Já a relação entre leitos_não_sus e leitos_sus é menor (0,781), valor

considerado moderado, sugerindo que essas variáveis podem fornecer informações distintas e foram mantidas no modelo.

Tabela 13. Correlação entre as variáveis que representam a quantidade leitos hospitalares

Variável 1	Variável 2	Correlação
leitos_total	leitos sus	0,960
leitos_total	leitos não sus	0,925
leitos_hab	leitos_sus_hab	0,906
leitos não sus	leitos sus	0,781

Fonte: Elaborado pelo autor

Na avaliação das informações relacionadas ao PIB municipal foi possível identificar uma alta correlação entre as variáveis PIB e Total geral e as demais, levando à exclusão dessas duas variáveis, conforme Tabela 14.

Tabela 14. Correlação entre as variáveis que representam o PIB municipal

Variável 1	Variável 2	Correlação
PIB (1)	Total geral	0,998
PIB (1)	Total (exclusive administração pública)	0,981
PIB (1)	Impostos	0,972
PIB (1)	Administração pública	0,916
PIB (1)	Indústria	0,846
Total geral	Total (exclusive administração pública)	0,978
Total geral	Impostos	0,958
Total geral	Administração pública	0,927

Fonte: Elaborado pelo autor

Por fim, na análise das receitas identificou-se uma forte correlação de 0,999 entre as receitas correntes e as receitas totais, indicando que a maior parte da arrecadação municipal é composta por receitas correntes. Com esse resultado as despesas totais foram retiradas do modelo. Já na análise das variáveis demográficas e estruturais, não foram encontradas correlações acima de 90%.

4.8 TREINAMENTO DOS MODELOS EXCLUINDO AS VARIÁVEIS ALTAMENTE CORRELACIONADAS

Um dos objetivos deste trabalho foi avaliar o impacto da exclusão de variáveis altamente correlacionadas no desempenho do modelo *Random Forest*, tanto em versão padrão quanto após a otimização de hiperparâmetros. O primeiro cenário, conforme apresentado na Tabela 15, mostra que a remoção de variáveis altamente correlacionadas praticamente não alterou o desempenho do modelo. As métricas MAE, MSE e R^2 permaneceram muito próximas, indicando que as variáveis excluídas provavelmente apresentavam redundância e não contribuíam com informações relevantes para a previsão.

Tabela 15. Comparação dos modelos padrão antes e após a exclusão das variáveis altamente correlacionadas

Métrica	Random Forest (Todas as Variáveis)	Random Forest (Excluindo variáveis altamente correlacionadas)
MAE	1.442.655,41	1.446.678,99
MSE	$87,59 \times 10^{12}$	$87,16 \times 10^{12}$
R²	0,85	0,85

Fonte: Elaborado pelo autor

Após a otimização dos hiperparâmetros, os modelos continuaram apresentando métricas muito próximas, mantendo a melhoria de desempenho em comparação com as versões padrão. A exclusão das variáveis altamente correlacionadas não comprometeu o desempenho, as diferenças nas métricas foram desprezíveis, conforme demonstrado na Tabela 16.

Tabela 16. Comparação dos modelos otimizados antes e após a exclusão das variáveis altamente correlacionadas

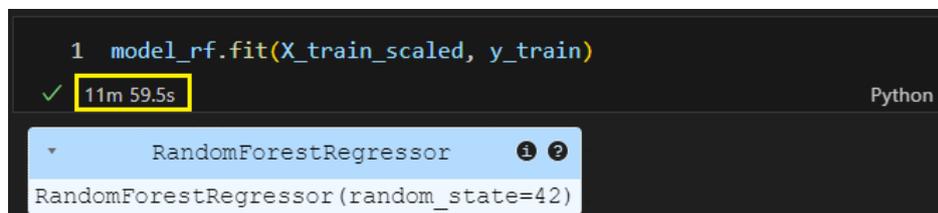
Métrica	Random Forest otimizado (Todas as Variáveis)	Random Forest otimizado (Excluindo variáveis altamente correlacionadas)
MAE	491.934,75	495.951,96
MSE	$10,95 \times 10^{12}$	$10,95 \times 10^{12}$

Métrica	Random Forest otimizado (Todas as Variáveis)	Random Forest otimizado (Excluindo variáveis altamente correlacionadas)
R ²	0,98	0,98

Fonte: Elaborado pelo autor

A exclusão de variáveis altamente correlacionadas não resultou em melhorias de desempenho no modelo *Random Forest*, mesmo após a otimização dos hiperparâmetros. No entanto, a remoção dessas variáveis redundantes contribuiu para a simplificação do modelo e para a redução no tempo de treinamento, conforme ilustrado na Figura 11.

Figura 11. Tempo de treinamento do modelo RF otimizado



```

1 model_rf.fit(X_train_scaled, y_train)
✓ 11m 59.5s Python
RandomForestRegressor ⓘ ?
RandomForestRegressor(random_state=42)

```

Fonte: Elaborado pelo autor

Observou-se uma redução no tempo de treinamento. O modelo original, com todas as variáveis, levou 17 minutos e 31 segundos para ser ajustado aos dados, enquanto o modelo sem as variáveis altamente correlacionadas foi treinado em 11 minutos e 59 segundos. Essa diferença de tempo, apesar de representar uma diminuição de 35%, não representa um ganho muito significativo ao considerar que o modelo será treinado apenas uma vez por ano na aplicação para o TCE-SP. Porém em situações em que os modelos serão retreinados dinamicamente com uma certa frequência, reduzir a necessidade computacional sem perda de desempenho pode ser relevante. Esse resultado demonstra que, para determinados casos, um conjunto de dados livre de redundâncias pode oferecer vantagens práticas.

5 PROPOSTA APLICADA

Este capítulo apresenta uma proposta de aplicação prática dos resultados alcançados para apoiar o trabalho de auditoria. O modelo treinado com os parâmetros otimizados tem como resultado uma tabela com os valores previstos no exercício em análise por cada uma das 644 prefeituras municipais, classificados por subelemento de despesa. O algoritmo RF prevê um valor para cada observação, com seus resultados exemplificados na Tabela 17.

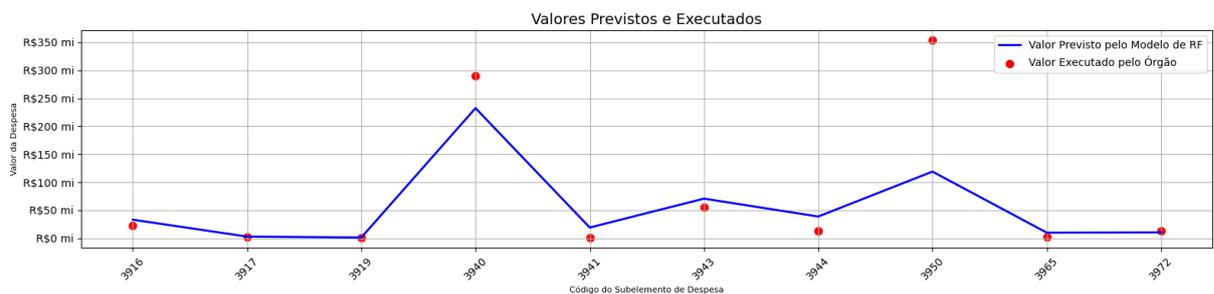
Tabela 17. Resultados previstos pelo modelo para uma Prefeitura Municipal

Exercício	Código IBGE	Código do Subelemento	Descrição do Subelemento	Valor Previsto pelo Modelo
2023	3509502	3916	Manutenção e conservação de bens imóveis	33.127.298,25
2023	3509502	3917	Manutenção e conservação de máquinas e equipamentos	3.001.366,66
2023	3509502	3919	Manutenção e conservação de veículos	1.306.016,95
2023	3509502	3940	Programa de alimentação do trabalhador	232.400.820,05
2023	3509502	3941	Fornecimento de alimentação	19.144.904,69
2023	3509502	3943	Serviços de energia elétrica	70.774.092,64
2023	3509502	3944	Serviços de água e esgoto	38.793.092,37
2023	3509502	3950	Serviço médico-hospitalar, odontológico e laboratorial	119.050.871,19
2023	3509502	3965	Serviços de apoio ao ensino	10.054.335,09
2023	3509502	3972	Vale-transporte	10.482.721,10

Fonte: Elaborado pelo autor

Com essas informações é possível gerar um gráfico com os valores previstos e aqueles efetivamente executados, conforme Figura 12.

Figura 12. Gráfico comparativo com os valores previstos pelo modelo e aqueles efetivamente executados



Fonte: Elaborado pelo autor

Dessa forma o auditor pode se basear na diferença entre a despesa prevista e a realizada pelo modelo e, usando seu julgamento profissional, decidir quais despesas devem ser examinadas com maiores detalhes. Nesse cenário o auditor poderia se basear na diferença absoluta ou percentual entre os valores previstos e realizados para escolher uma amostra.

Além das diferenças, para sistematizar ainda mais esse processo, propõe-se a utilização de um intervalo de confiança de 95% com o objetivo de destacar as despesas que efetivamente se distanciaram dos valores previstos. Para gerar o intervalo de confiança, é possível explorar a forma como o modelo RF gera as previsões, o valor previsto corresponde à média das previsões individuais geradas por múltiplas árvores de decisão. Além do valor final previsto pelo modelo, também é possível acessar as previsões geradas por cada uma das árvores. Esses valores podem ser utilizados para construir intervalos de confiança empíricos para cada registro. Com esses valores é possível calcular os percentis 2,5% e 97,5% da distribuição, formando um intervalo de confiança de 95%, como ilustrado na Figura 13.

Figura 13. Gráfico comparativo com os valores previstos pelo modelo e aqueles efetivamente executados



Fonte: Elaborado pelo autor

Nota: Além de incluir o intervalo de confiança, os pontos que representam os valores previstos foram coloridos em verde quando encontram-se dentro do intervalo de confiança e em vermelho quando estão fora dele.

É importante destacar que o intervalo de confiança gerado não é derivado de suposições clássicas de regressão. Trata-se de um intervalo empírico que reflete a incerteza real do processo preditivo decorrente das diferentes árvores de decisão geradas pelo modelo.

A partir desses pontos de atenção produzidos pela análise, o auditor pode verificar o detalhamento desses subelementos de despesa identificando o procedimento licitatório utilizado, as empresas contratadas e a descrição da despesa. De posse dessas informações, caso julgue necessário, pode escolher realizar procedimentos de auditoria para obter evidências sobre a exatidão, eficiência e legalidade dos gastos.

Para dar ainda mais celeridade à análise e permitir uma visualização intuitiva dos resultados, recomenda-se que eles sejam integrados a um painel interativo que permita que o usuário: (i) visualize os valores previstos pelo modelo e os intervalos de confiança esperados para cada subelemento e município, além das diferenças entre os valores previstos e os efetivamente liquidados; (ii) filtre na própria ferramenta os casos em que a despesa realizada está fora do intervalo de confiança; (iii) selecione cada um desses casos no painel de maneira que sejam exibidos os empenhos correspondentes, com detalhamento como modalidade de licitação, nome do credor, histórico do empenho, valores empenhados, liquidados e pagos; (iv) possibilite a exportação dos dados em formato csv ou xlsx para documentação em papéis de trabalho e utilização em relatórios.

6 DISCUSSÃO

A etapa inicial consistiu na estruturação do processamento dos dados. Para tanto, foram criados scripts na linguagem *Python* para acessar e consolidar os valores das despesas municipais, dos índices de inflação e demais bases necessárias à execução do projeto. Embora se trate de uma etapa bastante operacional, ela é fundamental para garantir a qualidade dos dados e agilidade na geração de análises em momentos futuros, pois, como destaca Ludermir (2021), a escolha das variáveis é apenas uma fase do processo de *machine learning*, sendo essencial contar com um conjunto de dados bem estruturado e representativo.

Com relação à despesa, optou-se por trabalhar com os dados de despesa liquidada, por refletir o momento em que a obrigação foi efetivamente reconhecida pelo órgão, conferindo um melhor ajuste temporal às informações. Adicionalmente, foi desenvolvido um *script* para automatizar a correção monetária das despesas, pois os valores disponíveis constavam nos bancos de dados em termos nominais. Para o cálculo da inflação acumulada, utilizou-se o IPCA mensal, abrangendo o período de janeiro de 2010 a dezembro de 2023. Esse fator foi aplicado a todas as despesas registradas com base na data de liquidação, uniformizando os dados em relação ao tempo e prevenindo uma interpretação equivocada pelo modelo.

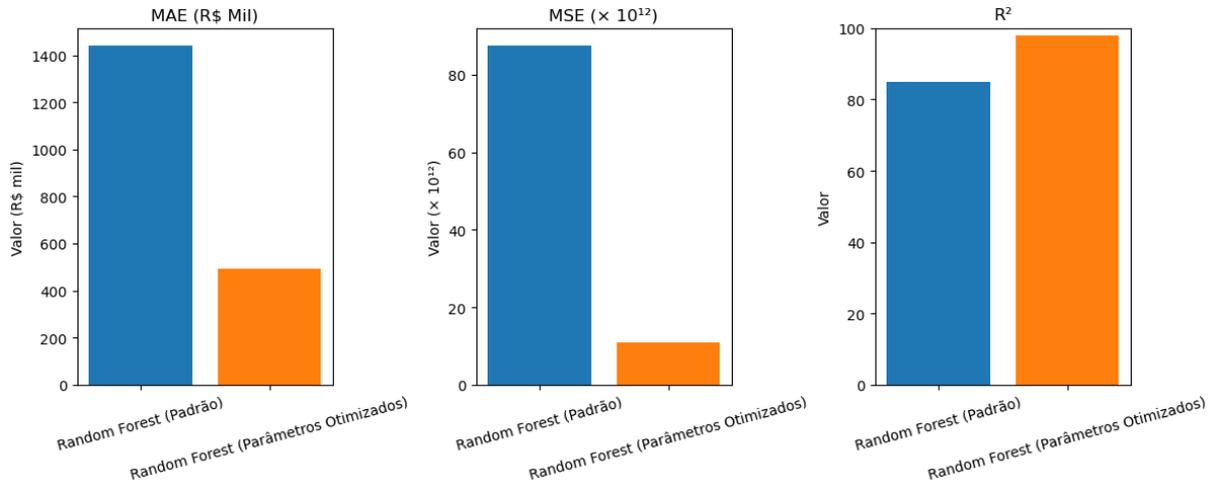
A partir das informações tratadas, foram treinados três modelos preditivos, RF, *LightGBM* e SVM e seus resultados foram analisados com base no erro médio absoluto (MAE), erro quadrático médio (MSE - *Mean Squared Error*) e coeficiente de determinação (R^2). Essa análise foi importante para avaliar qual deles se adapta melhor ao conjunto de dados disponíveis e aos objetivos do estudo.

Os resultados iniciais indicaram que o RF apresentou o melhor desempenho entre os modelos testados, alcançando um MAE e MSE menores e R^2 mais alto. O *LightGBM*, embora bastante eficiente em termos computacionais, demonstrou um poder preditivo inferior e erros maiores. Já o modelo SVM obteve o pior desempenho, com um R^2 negativo e tempo de treinamento extremamente elevado, o que demonstra que não é adequado aos dados disponíveis e foi desconsiderado nas etapas subsequentes.

Em seguida foi realizado o processo de otimização dos hiperparâmetros utilizando *GridSearchCV* da biblioteca *ScikitLearn* em sua versão 1.4.2 com validação cruzada e utilizando o MAE como critério para definição dos melhores parâmetros em cada modelo. Esse procedimento resultou em melhorias substanciais para o modelo RF com redução do MAE em

66%, do MSE em 87,5% e aumento do R^2 0,98, indicando um modelo altamente preciso, conforme Figura 14.

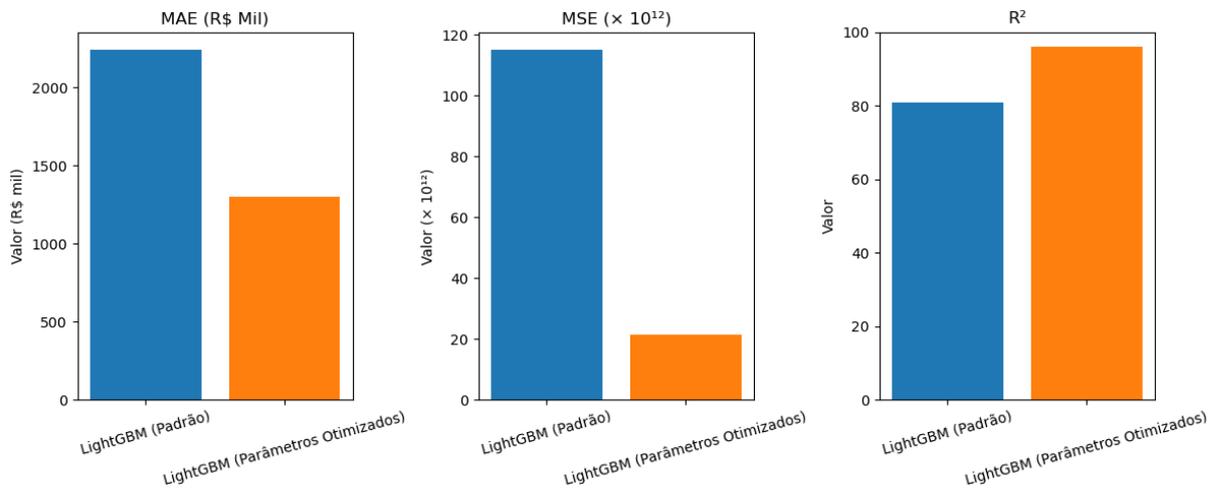
Figura 14. Resultado da otimização dos hiperparâmetros no modelo *Random Forest*



Fonte: Elaborado pelo autor

O modelo *LightGBM* também apresentou avanços significativos após a otimização, com um R^2 final de 0,96, embora ainda inferior à RF em relação aos erros. Os resultados do ajuste de seus parâmetros pode ser observado na Figura 15.

Figura 15. Resultado da otimização dos hiperparâmetros no modelo *LightGBM*



Fonte: Elaborado pelo autor

Esses resultados reforçam a importância da busca por configurações ideais de hiperparâmetros como etapa essencial para a melhoria do desempenho dos modelos preditivos.

A análise da importância das variáveis e da correlação entre elas evidenciou a presença de redundâncias, especialmente nas variáveis relacionadas à educação e saúde. Após a exclusão das variáveis altamente correlacionadas, as métricas MAE, MSE e R^2 se mantiveram praticamente inalteradas demonstrando que elas impactavam muito pouco o desempenho dos modelos. Apesar de não produzir uma variação significativa no poder preditivo do modelo, a remoção dessas variáveis resultou em uma redução de 35% no tempo de treinamento, demonstrando que a simplificação do modelo pode ser vantajosa, especialmente em contextos que exijam treinamentos frequentes ou quando houver pouca capacidade computacional disponível. Esse resultado reforça a importância de avaliar a correlação entre variáveis, não apenas em uma tentativa de aprimorar os modelos resultantes, mas também como estratégia para redução de custos computacionais.

Esse conjunto de resultados sustenta a conclusão de que a aprimoramento do pré-processamento, da seleção de variáveis e da calibração de hiperparâmetros melhora significativamente a performance de algoritmos de *machine learning*. Também foi possível identificar o modelo RF como o mais indicado para previsão das despesas com base nos dados utilizados nesse estudo. Sugere-se a adoção do modelo *Random Forest* gerado com hiperparâmetros otimizados para ferramentas de previsão de despesas municipais no âmbito do TCE-SP.

Embora a literatura não apresente modelos com o mesmo nível de detalhamento aplicado à despesa pública utilizada nesse estudo, evidências anteriores indicam que os modelos baseados em árvores de decisão, especialmente o RF, apresentam bom equilíbrio entre desempenho preditivo e robustez em aplicações com grande volume de dados, além de serem resistentes a *overfitting* (Zöllner & Huber, 2021; Montevechi et al., 2024; Silva et al., 2025).

Em análises financeiras realizadas no Brasil, como no estudo de dos Santos (2021), o RF também se destacou apresentando maior poder preditivo para dados heterogêneos, situação semelhante ao apresentado nessa pesquisa. De maneira complementar, de Oliveira Palmas (2023), criou modelos de ML a partir de dados da força aérea portuguesa, com objetivo de auxiliar o controle interno na identificação de desvios orçamentários. Nesse contexto os modelos baseados em árvore de decisão obtiveram os melhores resultados (de Oliveira Palmas, 2023).

Por fim, Raduenz (2020) realizou uma revisão bibliográfica selecionando nove artigos que utilizaram técnicas de ML para identificar *outliers*, fraudes e apontar desvios de custos em auditorias de contas médicas, constatando que o algoritmo de RF foi o mais aplicado nos trabalhos compilados, considerando seu desempenho eficaz.

7 CONCLUSÃO

O presente trabalho teve como objetivo principal aperfeiçoar o modelo utilizado pelo TCE-SP e propor uma abordagem sistematizada para sua aplicação no contexto da fiscalização dos recursos públicos, com o desenvolvimento de uma ferramenta intuitiva voltada para o uso dos auditores. O modelo de RF, com o acréscimo e tratamento de novas variáveis, ajuste de hiperparâmetros e exclusão das variáveis altamente correlacionadas, alcançou um erro absoluto médio (MAE) de R\$ 491.934,75 e um coeficiente de determinação (R^2) de 0,98, explicando com elevada precisão a variabilidade observada nos dados. Esses resultados demonstram um desempenho significativamente superior na previsão da despesa liquidada por subelemento nas prefeituras municipais paulistas, tanto com relação ao modelo anterior utilizado quanto aos novos modelos testados (SVM e LightGBM). Além disso, foram desenvolvidos intervalos de confiança empíricos permitindo identificar os desvios mais relevantes entre valores previstos e realizados, oferecendo subsídios objetivos para a seleção de amostras a serem examinadas no processo de auditoria. A aplicação dessa metodologia possibilita que auditores priorizem subelementos que destoam dos padrões previstos, promovendo uma seleção mais direcionada e eficiente das despesas a serem examinadas.

Apesar da melhoria significativa alcançada com a identificação de novas variáveis, automação da coleta e tratamento dos dados, este estudo enfrentou algumas limitações que devem ser destacadas. A principal delas refere-se à dificuldade de acesso a dados atualizados em tempo hábil. Como o processo de fiscalização demanda celeridade na disponibilização desse tipo de análise para que seja utilizada no processo de planejamento de auditoria, a indisponibilidade dos dados em tempo hábil compromete a utilidade prática do modelo, restringindo as variáveis que podem ser utilizadas. Além disso, aspectos institucionais e políticos que influenciam a execução orçamentária não foram considerados: mudanças na legislação, alternância no chefe do poder executivo e crises econômicas ou ambientais podem afetar a execução da despesa e por consequência podem ser relevantes para a capacidade preditiva do modelo.

Para trabalhos futuros, recomenda-se a incorporação de variáveis qualitativas e institucionais, para avaliar sua influência no modelo. Também é possível verificar se a utilização de regressões para estimar os valores de variáveis que não são disponibilizadas tempestivamente poderia gerar um modelo com melhores resultados. Outra sugestão diz respeito à heterogeneidade identificada nas variáveis utilizadas, sendo possível investigar se

agrupar as prefeituras municipais com características semelhantes e gerar modelos específicos para cada grupo pode gerar modelos preditivos com melhor desempenho. A criação de modelos mais específicos, relacionados a determinadas funções de governo, como saúde e educação, também é uma possível análise futura.

A principal contribuição deste estudo reside na proposta e validação de um modelo preditivo baseado em aprendizado de máquina como instrumento auxiliar no planejamento das auditorias no setor público. Utilizando variáveis socioeconômicas e orçamentárias, a pesquisa oferece uma nova abordagem no contexto do controle externo, propondo a utilização prática da ciência de dados na fiscalização de órgãos públicos. A geração de previsões baseadas em dados históricos e a definição de intervalos de confiança permitem a identificação racional de valores discrepantes, resultando em uma ferramenta que pode ser ou adaptada por outros tribunais de contas e demais órgãos de controle. Para contribuir com as pesquisas futuras, todo código desenvolvido para tratamento dos dados e treinamento dos modelos preditivos está disponível no *GitHub* e pode ser acessado no endereço <https://github.com/FabricioMacieira/ML-Despesas>.

REFERÊNCIAS

- Adhikari, P., & Gårseth-Nesbakk, L. (2016). Implementing public sector accruals in OECD member states: Major issues and challenges. *Accounting Forum*, 40(2), 125–142.
- Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). An overview of machine learning classification techniques. *BIO Web of Conferences*, 97, 133.
<https://doi.org/10.1051/bioconf/20249700133>
- Angelo da Silva, A. M., da Silva Vieira, R., & Duarte, A. J. M. (2015). Efficiency of municipal legislative chambers. *Economia*, 16(1), 60–75.
<https://doi.org/10.1016/j.econ.2015.03.001>
- Assis, E. da S., Simas, J. de L., & Amaral, R. G. do. (2021). UTILIZAÇÃO DA TÉCNICA DE MACHINE LEARNING PARA IDENTIFICAÇÃO DE FACIES GEOLÓGICAS ATRAVÉS DE PERFIS GEOFÍSICOS: UM RELATO DE EXPERIÊNCIA. In *Tecnologia da Informação e Comunicação: pesquisas em inovações tecnológicas* (pp. 343–351). Editora Científica Digital. <https://doi.org/10.37885/211106822>
- Aurélien Géron. (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems. In *O'Reilly Media*.
- Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing*, 08(04), 341–357.
<https://doi.org/10.4236/jdaip.2020.84020>
- Borges, W. G., Leroy, R. S. D., Carvalho, L. F., Lima, N. C., & Oliveira, J. M. (2020). Implicações da Inteligência Artificial na Auditoria Interna no Brasil: Análise sob a Percepção de Profissionais. *Sociedade, Contabilidade e Gestão*, 15(1), 23–40.
https://doi.org/10.21446/scg_ufrj.v0i0.25284
- Brasil. (1964). *Lei nº 4.320, de 17 de março de 1964: Estatui normas gerais de direito financeiro para elaboração e controle dos orçamentos e balanços da União, dos Estados, dos Municípios e do Distrito Federal*.
https://www.planalto.gov.br/ccivil_03/leis/l4320.htm
- Brasil. (1967). *Decreto-Lei nº 200, de 25 de fevereiro de 1967: Dispõe sobre a organização da Administração Federal*. https://www.planalto.gov.br/ccivil_03/decreto-lei/del0200.htm

- Brasil. (1988). *Constituição da República Federativa do Brasil de 1988*.
https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm
- Brasil. (1998). *Lei nº 9.717, de 27 de novembro de 1998: Dispõe sobre regras gerais para a organização e funcionamento dos regimes próprios de previdência social dos servidores públicos*. https://www.planalto.gov.br/ccivil_03/leis/19717.htm
- Brasil. (2000). *Lei complementar nº 101, de 4 de maio de 2000: Estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal*.
https://www.planalto.gov.br/ccivil_03/leis/lcp/lcp101.htm
- Brasil. (2001). *Portaria Interministerial nº 163, de 4 de maio de 2001: Dispõe sobre a padronização das classificações orçamentárias da receita e despesa pública*.
http://www.orcamentofederal.gov.br/orcamentos-anuais/orcamento-2015-2/arquivos%20portarias-sof/portaria-interm-163_2001_atualizada_2015_02set2015.pdf
- Brasil. (2016). *Lei nº 13.303, de 30 de junho de 2016: Dispõe sobre o estatuto jurídico da empresa pública, da sociedade de economia mista e de suas subsidiárias*.
https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/lei/113303.htm
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. In *Classification and Regression Trees*.
<https://doi.org/10.1201/9781315139470>
- Brown-Liburud, H., & Vasarhelyi, M. A. (2015). *Big data and audit evidence*.
- Carvalho, S. S. T. N. (2021). *Impacto da Inteligência Artificial na Atividade de Auditoria: Equacionando Gargalos nos Repasses da União para Entes Subnacionais*.
- Chakri, P., Pratap, S., Lakshay, & Gouda, S. K. (2023). An exploratory data analysis approach for analyzing financial accounting data using machine learning. *Decision Analytics Journal*, 7, 100212. <https://doi.org/10.1016/j.dajour.2023.100212>
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7. <https://doi.org/10.7717/PEERJ-CS.623>
- Chye Koh, H., & Kee Low, C. (2004). Going concern prediction using data mining techniques. *Managerial Auditing Journal*, 19(3), 462–476.
- Conselho Federal de Contabilidade. (2009). *NBCT TA 530: Amostragem em auditoria*.
<https://cfc.org.br/tecnica/normas-brasileiras-de-contabilidade/normas-de-auditoria-independente/>

- Conselho Federal de Contabilidade. (2016). *NBC TA 320 (R1) – Materialidade no planejamento e na execução da auditoria*. https://www.cfc.org.br/wp-content/uploads/2016/01/NBC_TA_320_R1.pdf
- Costa, E. G. da. (2021). *Gestão documental: proposta de um ambiente para a recuperação de documentos corporativos*.
- de Oliveira, D. B. da S. (2020). *Estimativa de Sobrevida de Pacientes com Glioblastoma por meio de Algoritmos Baseados em Random Forests*.
- de Oliveira Palmas, P. (2023). *Aplicação de Machine Learning na previsão de Desvios e Controlo orçamental na Força Aérea Portuguesa* (Master's thesis, Universidade de Lisboa (Portugal)).
- Donepudi, S., Nakka, R., Thota, K. K., Ajmeera, M., Praveen, S. P., & Sindhura, S. (2025). Enhancing Person-Centric Health Care for Diabetes Prediction: A Comparative Study of LightGBM, XGBoost, and Hybrid LIGB Model. In *Enabling Person-Centric Healthcare Using Ambient Assistive Technology, Volume 2: Personalized and Patient-Centric Healthcare Services in AAT* (pp. 127–155). Springer.
- De Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. (2023). The choice of scaling technique matters for classification performance. *Applied Soft Computing, 133*, 109924. <https://doi.org/10.1016/j.asoc.2022.109924>
- dos Santos, A. P., de Oliveira Menezes, E. C., & Raupp, F. M. (2022). Prioridade do gasto público: Análise das microrregiões de Paranaguá-PR e Baixo Vale Do Itajaí-SC. *Revista de Administração Unimep, v. 19, n. 10*, 20–40.
- Efron, B. (2020). Prediction, Estimation, and Attribution. *Journal of the American Statistical Association, 115*(530). <https://doi.org/10.1080/01621459.2020.1762613>
- Florêncio, C. C. A. (2022). *Contabilidade Pública*.
- Francis, J. R. (2024). What exactly do we mean by audit quality? *Accounting in Europe, 21*(2), 123–133. <https://doi.org/10.1080/17449480.2023.2247410>
- Gama, J. R. (2012). *Contabilidade Pública I*.
- Gigio, R. B. (2024). *Desenvolvimento de uma plataforma para análise financeira e de sentimento no setor logístico: aplicações de machine learning e visualização interativa*.
- Hardiansyah, A., Modjo, A. S., & Yuli Vandika, A. (2024). APPLICATIONS OF MACHINE LEARNING IN DIAGNOSTICS AND SPECIAL EDUCATION INTERVENTIONS. In *Indonesian Journal of Education (INJOE)* (Vol. 4, Issue 1).
- Hutter, F., Kotthof, L., & Vanschoren, J. (2020). Automated machine learning: Methods, Systems, Challenges. In *The Springer Series on Challenges in Machine Learning*.

- Instituto Rui Barbosa. (2023). *NBASP 200: Princípios de auditoria financeira*.
https://nbasp.irbcontas.org.br/wp-content/uploads/jet-form-builder/78f1893678afbeaa90b1fa01b9cfb860/2023/11/NBASP-200_PRINCIPIOS-DE-AUDITORIA-FINANCEIRA-2.pdf
- Janitza, S., Strobl, C., & Boulesteix, A. L. (2013). An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics*, *14*.
<https://doi.org/10.1186/1471-2105-14-119>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017-December.
- Kluyver, T., Ragan-Kelley, B., Perez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & Jupyter Development Team. (2016). Jupyter Notebooks a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press. <https://doi.org/10.3233/978-1-61499-649-1-87>
- Law, K. K. F., & Shen, M. (2024). How does artificial intelligence shape audit firms? *Management Science*.
- Lazaretti, L. R., & Fochezatto, A. (2020). *Despesas municipais e tamanho ótimo da população: uma abordagem de painel dinâmico espacial para os municípios do Rio Grande do Sul*.
- Leitão, J. M., & Dantas, J. A. (2016). Materialidade em auditoria financeira no setor público: a prática internacional e a opinião de especialistas brasileiros. *Contabilidade, Gestão e Governança*, v. 19, 146–166.
- Leite, M. F. (2024). *Análise descritiva e exploratória sobre os dados de projetos de pesquisa e extensão disponibilizados no portal de dados abertos do IFRN*.
- Leite, R. R., & Werneck, L. M. T. (2012). *A POLÍTICA DE FUNDOS CONTÁBEIS: O QUE MUDA NA EDUCAÇÃO NOS MUNICÍPIOS GOIANOS*.
- Li, S., Dong, X., Ma, D., Dang, B., Zang, H., & Gong, Y. (2024). *Utilizing the LightGBM Algorithm for Operator User Credit Assessment Research*.
<https://doi.org/10.54254/2755-2721/75/20240503>
- Ludermir, T. B. (2021). Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. *Estudos Avançados*, *35*(101), 85–94. <https://doi.org/10.1590/s0103-4014.2021.35101.007>

- McKinney, W. (2018). *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. Novatec Editora.
- Ministério da Economia. (2021). *Portaria Conjunta STN/SOF/ME nº 103, de 5 de outubro de 2021: Estabelece regras sobre codificação de natureza de despesa*.
<https://www.in.gov.br/en/web/dou/-/portaria-conjunta-stn/sof/me-n-103-de-5-de-outubro-de-2021-347165769>
- Mishra, D. (2021). Regression: {An} {Explanation} of {Regression} {Metrics} {And} {What} {Can} {Go} {Wrong}. In *Medium*.
- Mitan, J. (2024). *Enhancing Audit Quality through Artificial Intelligence: An External Auditing Perspective*. <https://scholarworks.uark.edu/acctuht>
- Montevechi, A. A., Miranda, R. de C., Medeiros, A. L., & Montevechi, J. A. B. (2024). Advancing credit risk modelling with Machine Learning: A comprehensive review of the state-of-the-art. *Engineering Applications of Artificial Intelligence*, 137, 109082.
<https://doi.org/10.1016/j.engappai.2024.109082>
- Moraes, D. P. de. (2006). *Arrecadação tributária municipal: esforço fiscal, transferências e Lei de Responsabilidade Fiscal*.
- Nascimento, C. L. (2022). *O controle interno preventivo à luz do Sistema Alice: propostas de trilhas para detecção de anomalias na execução de programas sociais*. Fundação Getúlio Vargas.
- Nascimento, J. L. R. (2022). *Índice de priorização de objetos de auditoria: um estudo de caso para municípios sergipanos*.
- Neves, F. R., da Silva, P. B., & de Carvalho, H. L. M. (2019). Artificial ladies against corruption: searching for legitimacy at the Brazilian Supreme Audit Institution. *Revista de Contabilidade e Organizações*, 13, 31–50. <https://doi.org/10.11606/issn.1982-6486.rco.2019.158530>
- Njungle, N. B., Jahns, E., Wu, Z., Mastromauro, L., Stojkov, M., & Kinsy, M. A. (2025). GuardianML: Anatomy of Privacy-Preserving Machine Learning Techniques and Frameworks. *IEEE Access*, 13, 61483–61510.
<https://doi.org/10.1109/ACCESS.2025.3557228>
- Nunes, P. de C., Bellas, H., Paulino, É. T., Ramos, A., & Jatobá, A. (2024). Manutenção dos serviços de média e alta complexidade em contexto de alta transição de usuários: um estudo ecológico de séries temporais no Estado do Rio de Janeiro, Brasil. *Ciência & Saúde Coletiva*, 29(1). <https://doi.org/10.1590/1413-81232024291.16542022>

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12.
- Pimentel, J., Oliveira, G., Silva, M., Seufitelli, D., & Moro, M. (2021). Ciência de Dados com Reprodutibilidade usando Jupyter. In *Jornada de Atualização em Informática 2021*. <https://doi.org/10.5753/sbc.6757.3.1>
- Raduenz, J. C. Machine learning na Auditoria de contas médicas.
- Rahman, M. M., Usman, O. L., Muniyandi, R. C., Sahran, S., Mohamed, S., & Razak, R. A. (2020). A review of machine learning methods of feature selection and classification for autism spectrum disorder. In *Brain Sciences* (Vol. 10, Issue 12). <https://doi.org/10.3390/brainsci10120949>
- Rosa, C. S. (2018). *Estudo sobre as técnicas e métodos de análise de dados no contexto de Big Data*.
- Sá Filho, C. S. (2020). *Análise do risco fiscal da dívida pública nos estados brasileiros pelo exame do comportamento das despesas orçamentárias no período de 2008 a 2016*.
- Sales, L. J. (2017). *Proposta de modelo de classificação do risco de contratos públicos*.
- Santos, A. S. dos. (2021). *Previsão de insolvência corporativa: uma análise de empresas brasileiras de capital aberto por meio de aprendizado de máquina*.
- Santos, J. F. dos. (2024). *Astronomia hoje: um roteiro para caracterização de exoplanetas descobertos na missão TESS usando a linguagem Python*.
- São Paulo. (1989). *Constituição do Estado de São Paulo de 1989*. <https://www.al.sp.gov.br/repositorio/legislacao/constituicao-estadual/constituicao-estadual-1989.pdf>
- Savério, N. A. (2023). *Uso de Inteligência Artificial (IA) na Administração Pública Brasileira*.
- Scikit-learn developers. (2023). *Permutation Feature Importance*.
- Secretaria da Fazenda do Estado de São Paulo. (2024). *Consulta de despesa anual – Orçamento estadual*. <https://www.fazenda.sp.gov.br/SigeoLei131/Paginas/ConsultaDespesaAno.aspx?orgao=>
- Silva, R. C. C. da, Clímaco, F. G. N., Chagas, F. B., & Placido, K. F. (2025). Modelagem preditiva para sucesso acadêmico: um estudo de caso em um curso de ciência da computação. *REVISTA DELOS*, 18(63), e3586. <https://doi.org/10.55905/rdelosv18.n63-077>

- Silveira, M. R. (2022). Circulação, transportes e logística no Brasil. *Revista Da ANPEGE*.
<https://doi.org/10.5418/ra2022.v18i36.16202>
- Stanišić, N., Radojević, T., & Stanić, N. (2019). Predicting the type of auditor opinion: Statistics, machine learning, or a combination of the two? *The European Journal of Applied Economics*, 16(2). <https://doi.org/10.5937/ejae16-21832>
- Stock, A. (2025). Choosing blocks for spatial cross-validation: lessons from a marine remote sensing case study. *Frontiers in Remote Sensing*, 6.
<https://doi.org/10.3389/frsen.2025.1531097>
- Teshale, A. B., Htun, H. L., Vered, M., Owen, A. J., Ryan, J., Polkinghorne, K. R., Kilkenny, M. F., Tonkin, A., & Freak-Poli, R. (2025). Integrating Social Determinants of Health and Established Risk Factors to Predict Cardiovascular Disease Risk Among Healthy Older Adults. *Journal of the American Geriatrics Society*.
<https://doi.org/10.1111/jgs.19440>
- Tesouro Nacional. (2024). *Manual de contabilidade aplicada ao setor público (MCASP) – 2024*. <https://www.tesourotransparente.gov.br/publicacoes/manual-de-contabilidade-aplicada-ao-setor-publico-mcasp/2024/26>
- Tiossi, F. M., & Medeiros, P. C. (2024). *O TRIBUNAL DE CONTAS COMO ÓRGÃO DE CONTROLE EXTERNO*.
- Tojeiro, C. A. C. (2024). *Identificação de portscan com machine learning: uma análise dos impactos de algoritmos de oversampling e undersampling no desenvolvimento de sistemas de detecção de intrusão*.
- Tribunal de Contas do Estado de São Paulo. (2022). *Plano estratégico 2022–2026*.
<https://www.tce.sp.gov.br/gestao-estrategica/planejamento-estrategico-2022-2026>
- Tribunal de Contas do Estado de São Paulo. (2023a). *Documentação técnica do sistema Audesp: Plano de contas e tabelas auxiliares*.
<https://www.tce.sp.gov.br/audesp/documentacao>
- Tribunal de Contas do Estado de São Paulo. (2023b). *Regimento Interno*.
<https://www.tce.sp.gov.br/legislacao/lei-organica-e-regimento-interno/regimento-interno>
- Tribunal de Contas do Estado de São Paulo. (2024a). *Portal da transparência municipal - Conjunto de dados*. <https://transparencia.tce.sp.gov.br/conjunto-de-dados>
- Tribunal de Contas do Estado de São Paulo. (2024b). *Quadro de pessoal – Dezembro de 2024*.
https://www.tce.sp.gov.br/sites/default/files/transparencia/Quadro_Pessoal%20Dezembro%202024.pdf

- Ucoglu, D. (2020). Current machine learning applications in accounting and auditing. *Pressacademia*, 12(1). <https://doi.org/10.17261/pressacademia.2020.1337>
- Uludag, K. (2024). *Hyperparameters and Tuning Methods for Random Forest Using Python Sklearn Package Relevant to Psychology Studies* (pp. 207–222). <https://doi.org/10.4018/979-8-3693-2703-6.ch011>
- Wies, C., Miltenberger, R., Grieser, G., & Jahn-Eimermacher, A. (2023). Exploring the variable importance in random forests under correlations: a general concept applied to donor organ quality in post-transplant survival. *BMC Medical Research Methodology*, 23(1). <https://doi.org/10.1186/s12874-023-02023-2>
- Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1). <https://doi.org/10.3354/cr030079>
- Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., Ma, C., Yan, J., & Wang, X. (2021). LightGBM: accelerated genomically designed crop breeding through ensemble learning. *Genome Biology*, 22, 1–24.
- Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168(2). <https://doi.org/10.1088/1742-6596/1168/2/022022>
- Zaniboni, N. C., & Montini, A. de A. (2019). *Aplicações de Machine Learning em Contabilidade: Uma Revisão da Literatura Brasileira*.
- Zhang, Y., Gao, S., Cai, P., Lei, Z., & Wang, Y. (2023). Information entropy-based differential evolution with extremely randomized trees and LightGBM for protein structural class prediction. *Applied Soft Computing*, 136. <https://doi.org/10.1016/j.asoc.2023.110064>
- Zhao, X., Liu, Y., & Zhao, Q. (2024). Improved LightGBM for extremely imbalanced data and application to credit card fraud detection. *IEEE Access*.
- Zöllner, M.-A., & Huber, M. F. (2021). Benchmark and Survey of Automated Machine Learning Frameworks. In *Journal of Artificial Intelligence Research* (Vol. 70).