

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA E DE ARTES
GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO



**CIÊNCIA DE DADOS APLICADA NA DETECÇÃO DE FRAUDE EM
DEMONSTRAÇÕES FINANCEIRAS**

JOÃO PAULO TEODORO PEREIRA

GOIÂNIA
2025

JOÃO PAULO TEODORO PEREIRA

**CIÊNCIA DE DADOS APLICADA NA DETECÇÃO DE FRAUDE EM
DEMONSTRAÇÕES FINANCEIRAS**

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

Orientador:

Professor Me. André Luiz Alves

Banca examinadora:

Prof. Me. Fabricio Schlag

Prof.^a Esp. Nágela Bitar Lôbo

GOIÂNIA
2025

JOÃO PAULO TEODORO PEREIRA

**CIÊNCIA DE DADOS APLICADA NA DETECÇÃO DE FRAUDE EM
DEMONSTRAÇÕES FINANCEIRAS**

Trabalho de Conclusão de Curso aprovado em sua forma final pela Escola Politécnica e de Artes, da Pontifícia Universidade Católica de Goiás, para obtenção do título de Bacharel em Engenharia de Computação, em: 01/12/2025.

Orientador: Prof. Me. André Luiz Alves

Prof. Me. Fabricio Schlag

Prof.^a Esp. Nágela Bitar Lôbo

GOIÂNIA
2025

RESUMO

Este trabalho busca compreender a aplicação da Ciência de Dados e da lei de Benford como instrumentos auxiliares em auditoria de transações financeiras. Como estudo de caso, propõe-se analisar a conformidade dos repasses extraorçamentários do Estado de Goiás, referentes ao exercício de 2024, disponibilizados no portal da transparência do Estado de Goiás, com objetivo de verificar se as transações estão livres de distorções que indicam possíveis irregularidades, por meio da aplicação da lei de Benford, suportada por testes estatísticos. A metodologia adota o processo de Descoberta de Conhecimento em Bases de Dados do inglês *Knowledge Discovery in Databases* (KDD), abrangendo a coleta, limpeza, mineração de dados e validação com testes estatísticos Qui-Quadrado e Z-score, por meio de implementação em linguagem *Python*. Os resultados demonstram divergências significativa entre a frequência observada e a esperada pela lei de Benford, destacando-se uma quantidade além do esperado no primeiro dígito 1 e desvios compensatórios em outros dígitos, levando à rejeição da hipótese nula de conformidade. Conclui-se que a ferramenta computacional desenvolvida atua de forma eficaz como um filtro inicial para a detecção de padrões anômalos em auditorias, permitindo ao auditor direcionar o foco para transações de maior risco e otimizar os recursos de fiscalização.

Palavras-chave: Mineração de Dados. Auditoria Contábil. Lei de Newcomb-Benford. Testes Estatísticos.

ABSTRACT

This study aims to understand the application of Data Science and Benford's law as auxiliary instruments in the auditing of financial transactions. As a case study, the conformity of extra-budgetary disbursements of the State of Goiás for the fiscal year 2024, available on the State of Goiás Transparency Portal, is analyzed. The objective is to verify if the transactions are free from distortions indicating possible irregularities, through the application of Benford's Law supported by statistical tests. The methodology adopts the Knowledge Discovery in Databases (KDD) process, covering data collection, cleaning, mining, and validation using Chi-Square and Z-score statistical tests, implemented in Python. The results demonstrate significant divergences between the observed frequency and that expected by Benford's Law, highlighting a higher-than-expected occurrence of the first digit 1 and compensatory deviations in other digits, leading to the rejection of the null hypothesis of conformity. It is concluded that the developed computational tool acts effectively as an initial filter for detecting anomalous patterns in audits, allowing the auditor to focus on higher-risk transactions and optimize inspection resources.

Keywords: Data Mining. Accounting Auditing. Newcomb–Benford Law. Statistical Tests.

LISTA DE QUADROS

Quadro 1 - Diagrama de Venn aplicado a ciência de dados	13
Quadro 2 - Etapas do processo de KDD	14
Quadro 3 - Triângulo da fraude	17
Quadro 4 - Distribuição de probabilidade para o primeiro dígito	20
Quadro 5 - Etapas adotadas no estudo de caso	25
Quadro 6 - Portal da transparência - Gastos Governamentais.....	26
Quadro 7 - Dados extraídos e organizados.....	26
Quadro 8 - Resultado da extração do primeiro e segundo dígito	31
Quadro 9 - Comparação entre distribuição observada e esperada	33
Quadro 10 - Gráficos dos desvios	33
Quadro 11 - Contribuição de cada dígito para a estatística.....	35
Quadro 12 - Gráfico de z-scores por dígito	36

LISTA DE TABELAS

Tabela 1 - Probabilidades esperadas dos dígitos de acordo com a lei de Benford ...	21
Tabela 2 - Amostra dos cinco primeiros registros do arquivo após ajustes	28
Tabela 3- exemplo de registros removidos	30
Tabela 4 - Registro negativos que foi convertido em positivo.....	30
Tabela 5 - Distribuição observada e esperada do primeiro dígito.....	32

LISTA DE SIGLAS

CFC	Conselho Federal de Contabilidade
KDD	<i>Knowledge Discovery in Databases</i>
NBC	Normas Brasileiras de Contabilidade
NBC TA	Normas Brasileiras de Contabilidade Técnica de Auditoria

SUMÁRIO

1. INTRODUÇÃO	10
2. REFERENCIAL TEÓRICO.....	12
2.1. Ciência de dados	12
2.2 Mineração de dados	13
<i>2.2.1 Tarefas da Mineração de Dados.....</i>	<i>15</i>
<i>2.2.2. Aplicações da mineração de dados.....</i>	<i>16</i>
2.3 Fraude nas demonstrações financeiras	16
2.4. Mineração de dados na detecção de fraudes em demonstrações financeiras	18
2.5. A lei de Newcomb-Benford	18
<i>2.5.1. Aplicações em auditoria e detecção de fraudes</i>	<i>21</i>
<i>2.5.2. Testes Estatísticos de Conformidade</i>	<i>22</i>
3. METODOLOGIA	24
3.1. Coleta de Dados	25
3.2. Apresentação e validação dos dados.....	27
3.3. Limpeza dos Dados.....	29
3.4. Aplicação da lei de Benford.....	31
3.5. Testes estatísticos.....	34
<i>3.5.1. Teste de aderência Qui-Quadrado (χ^2).....</i>	<i>34</i>
<i>3.5.2. Teste Z-score.....</i>	<i>35</i>
4. RESULTADOS	37
4.1. Análise da distribuição de frequências	37
4.2. Análise do teste Qui-Quadrado	38
4.3. Análise do teste Z-score	38
4.4. Síntese dos resultados	39
5. CONCLUSÃO	40
REFERÊNCIAS.....	42

1. INTRODUÇÃO

A auditoria das transações contábeis tem se tornado cada vez mais desafiadora, diante do grande volume de dados gerados e da complexidade das operações financeiras. Neste cenário a Ciência de Dados se mostra uma importante ferramenta, capaz de auxiliar na identificação de padrões em grande base de dados e apoiar a tomada de decisão dos profissionais de auditoria. Sua utilização contribui com avanços significativos na detecção de fraudes financeiras, visto que permite a identificação de distorções relevantes que muitas vezes passam despercebidas quando utilizado apenas métodos tradicionais de auditoria.

Um obstáculo enfrentado por auditores é a grande quantidade de transações a serem analisadas, onde normalmente dependem de amostragem em vez da base integral, e a dificuldade em integrar dados de diferentes fontes, isso associado a crescente adaptação e sofisticação de esquemas fraudulentos dificultam a identificações de fraude.

Diante desses desafios, a Ciência de Dados contribui com princípios e técnicas de análise de dados, possibilitando o desenvolvimento de ferramentas computacionais analíticas e eficientes para o uso em auditoria financeira.

A aplicação da lei de Benford é uma ferramenta importante para análise exploratória e detecção de anomalias. Observado inicialmente por Simon Newcomb (1881) e formalizada posteriormente por Frank Benford (1938), é ideal para a auditoria, pois dados financeiros reais tendem a seguir a distribuição logarítmica apresentada pelos autores, enquanto números manipulados intencionalmente se desviam dela, atuando como um "sinalizador" (*red flag*) eficiente, capaz de identificar anomalias que justificam uma investigação mais minuciosa por parte do profissional de auditoria.

Neste trabalho foi realizado um estudo de caso, para análise de conformidade dos desembolsos extraorçamentários do Estado de Goiás, no exercício de 2024, com aplicação da lei de Benford, utilizando o processo de Descoberta de Conhecimento em Base de Dados, do inglês *Knowledge Discovery in Databases* (KDD) como metodologia, no intuito de identificar se houve ou não, anomalias estatísticas indicativas de distorções por manipulação de valores. Com os resultados obtidos

demonstra-se que a aplicação de conceitos da ciência de dados e mineração de dados contribuem com melhorias significativas para auditoria de movimentações financeiras.

Para compreender essas melhorias na prática, é necessário primeiramente estabelecer as bases conceituais que sustentam este trabalho. O capítulo a seguir apresenta o referencial teórico, onde é abordado os fundamentos da Ciência de Dados, Mineração de Dados, fraudes financeiras e a definição matemática da lei de Newcomb-Benford, com sua aplicação na detecção de fraudes.

2. REFERENCIAL TEÓRICO

Este capítulo fundamenta os conceitos teóricos relevantes para a compreensão e desenvolvimento do projeto, sendo: princípios da ciência de dados e o processo de descoberta de conhecimento em bases de dados, com destaque para as técnicas de mineração de dados; contextualização de fraude em demonstrações financeiras e a teoria do triângulo da fraude; A lei de Benford e sua aplicação como uma ferramenta computacional em auditorias contábeis.

2.1. Ciência de dados

Provost e Fawcett (2013) definem ciência de dados de forma ampla, sendo um conjunto de princípios fundamentais que guiam a extração do conhecimento a partir dos dados, dando suporte ao pensamento analítico e contribuindo para a tomada de decisão estratégica.

Apesar de ser confundida com mineração de dados, a ciência de dados vai além da utilização de algoritmos e tecnologias de mineração. Ela se relaciona com várias áreas do conhecimento, em especial a matemática e estatística, sendo estas áreas fundamentais para a prática de ciência de dados (PROVOST e FAWCETT, 2013). Os autores ainda sustentam que a ciência de dados requer um equilíbrio entre processos estruturados e a criatividade na solução de problema, visto que metodologias rígidas podem restringir a exploração e a inovação.

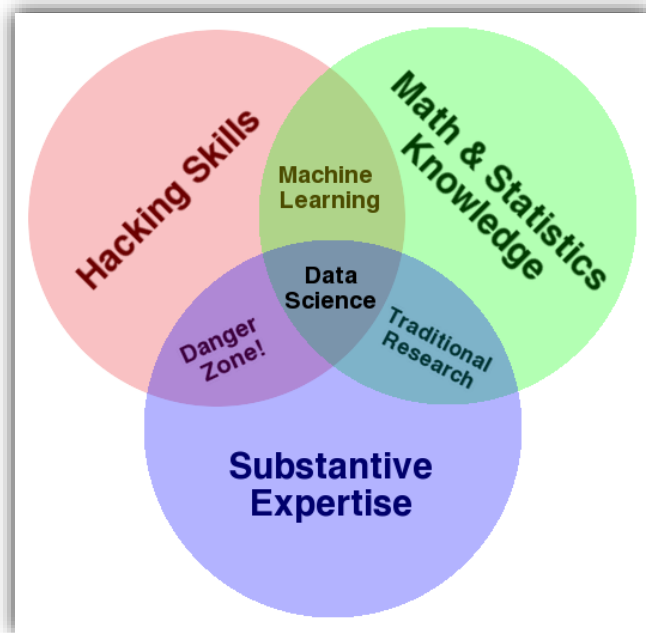
Para Grus (2016) a ciência de dados se encontra na interseção entre três áreas fundamentais:

1. **Habilidades hacker:** refere-se à habilidade de programar, na capacidade de manipular dados, criar algoritmos e processar informações brutas;
2. **conhecimentos de estatística e matemática:** a base necessária para extrair informações relevantes, aplicar testes de hipótese e modelar fenômenos, garantindo que além de processados, os dados sejam compreendidos analiticamente;
3. **competências significativas:** o entendimento do domínio do negócio, que permite interpretar os resultados corretamente.

Sendo que a matemática e estatística são a base para a prática de ciência de dados.

O Quadro 1, proposta por Drew Conway, ilustra essa integração. O diagrama demonstra que a Ciência de Dados ocorre no centro, onde as três áreas se encontram.

Quadro 1 - Diagrama de Venn aplicado a ciência de dados



Fonte: CONWAY, Drew. The Data Science Venn Diagram (2013)

É importante notar no diagrama a presença da "Zona de Perigo" (*Danger Zone*), que ocorre na interseção entre habilidades de programação (*Hacking Skills*) e conhecimento do negócio (*Substantive Expertise*), mas deixa de lado os conhecimentos de matemática e estatística. Desta forma, é possível gerar análises que parecem corretas, mas podem ser inválidas estatisticamente. Isso reforça a importância dos testes estatísticos.

2.2 Mineração de dados

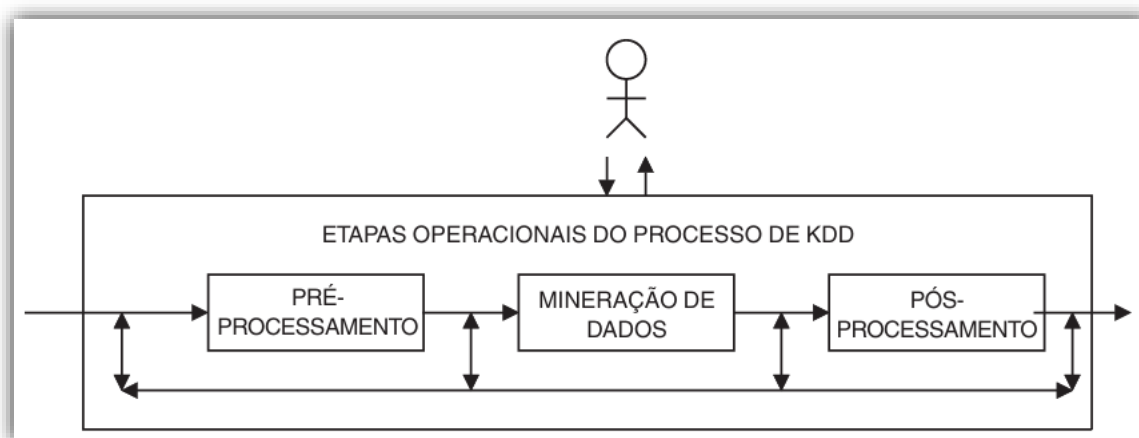
Segundo Silva, Peres e Boscarioli (2016) a mineração de dados é o esforço de extrair padrões e informações relevantes de grandes volumes de dados, de forma automática ou semiautomática, por meio da exploração analítica, e tem com objetivo de dar suporte à geração de conhecimento útil para a tomada de decisão. É uma área interdisciplinar utilizando de conceitos de computação, estatística, inteligência artificial, aprendizado de máquina e banco de dados.

A mineração de dados está intimamente relacionada ao processo de KDD sendo uma das etapas deste processo (GOLDSHMIDT e PASSOS, 2005; SILVA, PERES e BOSCARIOLI. 2016).

O KDD surge da necessidade de ferramentas que auxiliam de forma automática e inteligente na análise, interpretação e relacionamento entre os dados em grandes e múltiplas base de dados, possibilitando a tomada de decisões estratégicas para o contexto da aplicação. É um processo dividido em várias etapas, iterativas, interativas e não trivial que tem por objetivo a identificação de novos padrões intrínsecos, compreensíveis e úteis a partir de grandes conjuntos de dados (GOLDSHMIDT e PASSOS, 2005; SILVA et al. 2016).

Conforme demonstrado no Quadro 2, Goldshmidt e Passos (2005) dividem o processo de KDD em 3 etapa operacionais, sendo a mineração de dados a etapa central, onde algoritmos são aplicados para extrair padrões.

Quadro 2 - Etapas do processo de KDD



Fonte: Goldschmidt e Passos (2005)

1. **Pré-processamento:** É a etapa de captação e preparação dos dados, e envolve tarefas essenciais para garantir a qualidade dos dados e influencia diretamente o resultado das etapas seguintes. Silva et al (2016) define como funções da etapa de pré-processamento:
 - a. **Seleção** - escolha dos dados e atributos relevantes;
 - b. **Limpeza** - tratamento de valores nulos, inconsistência, correção de ruídos e *outliers*;
 - c. **Integração** - consolidação de dados de fontes distintas, em um único conjunto de dados;
 - d. **Transformação** - normalização e padronização dos dados.

2. **Mineração de Dados:** é a etapa central onde os algoritmos computacionais são alimentados com os dados pré-processados, para identificar padrões e extrair conhecimento.
3. **Pós-processamento:** nesta etapa é realizado o tratamento dos resultados obtidos na mineração, onde, os novos padrões gerados são avaliados quanto à sua utilidade. O conhecimento adquirido é interpretado, com apoio de especialistas do domínio, e apresentado de forma compreensível, como relatórios ou visualização gráfica, servindo para apoiar a tomada de decisão.

2.2.1 Tarefas da Mineração de Dados

As tarefas da mineração de dados podem ser divididas em duas categorias, sendo, preditiva e descritiva. As tarefas preditivas utilizam dados para prever valores futuros ou desconhecidos, enquanto as tarefas descritivas buscam encontrar padrões que descrevam os dados de forma que possa ser interpretado pelo ser humano (SILVA et al. 2016). Dentro dessas categorias, destacam-se as tarefas de:

1. **Predição:** São tarefas supervisionadas, como a classificação e regressão. A classificação busca descobrir uma função, ou modelo capaz de mapear um registro de dados em uma classe (categoria) predefinida. A regressão, apesar de similar à classificação, se aplica quando o valor a ser previsto é numérico e contínuo (GOLDSCHMIDT e PASSOS, 2005).
2. **Agrupamento:** O Agrupamento (ou Clusterização) é uma tarefa não supervisionada que organiza os dados em grupos (*clusters*) com base em suas similaridades. Desta forma espera-se que elementos dentro de um mesmo grupo sejam similares entre si, ao passo que são distintos dos elementos de outros grupos. Esta técnica é frequentemente utilizada por exemplo, para a segmentação de clientes, identificando perfis de comportamento sem se ter um conhecimento prévio das classes (GOLDSCHMIDT e PASSOS, 2005).
3. **Descoberta de associação:** Esta tarefa tem por objetivo identificar a ocorrência frequente e simultânea de itens em um conjunto de dados. Um exemplo de aplicação é na análise de cesta de compras, para identificar produtos que frequentemente são comprados em conjunto. (GOLDSCHMIDT e PASSOS, 2005).

Segundo os autores essas são apenas algumas tarefas da mineração de dados, existem ainda outras relevantes, como a sumarização, que descreve características comuns de um conjunto de dados, e a detecção de desvios (*outliers*), que tem como objetivo identificar elementos do conjunto que fogem da distribuição padrão esperada, sendo muito utilizada para a detecção de alguns tipos de fraudes.

2.2.2. Aplicações da mineração de dados

As aplicações da mineração de dados abrangem diversas áreas como computação, administração e engenharia, oferecendo suporte a decisões estratégicas baseadas em evidências. Como exemplo a análise de dados transacionais, voltado para a identificação de padrões de consumo.

No varejo, é utilizada para estratégias de marketing através da identificação de padrões em cestas de compras. Na medicina, como auxílio no diagnóstico. (SILVA et al., 2016).

Na área financeira, é uma importante ferramenta para a análise de concessão de crédito e detecção de fraudes em cartões.

Além disso, destaca-se sua importância no combate a fraudes, por meio das tarefas da mineração, aplicadas em análise de cartões de crédito, seguros e na área fiscal relacionada a arrecadação de tributos, onde padrões anômalos indicam possíveis comportamentos fraudulentos (GOLDSCHMIDT e PASSOS, 2005).

Em resumo, o potencial de aplicação da mineração de dados é amplo, podendo ser utilizado em qualquer domínio que contenha volumes razoáveis de dados históricos.

2.3 Fraude nas demonstrações financeiras

De acordo com as Normas Brasileiras de Contabilidade Técnica de Auditoria (NBC TA) 240 (Conselho Federal de Contabilidade (CFC), 2016), fraude é um ato intencional cometido por um ou mais indivíduos, responsáveis pela governança, empregados ou terceiros, com objetivo de obter vantagem injusta ou ilegal. A norma afirma que fraude não se confunde com erro, que apesar de também corresponder a uma distorção, este não é intencional, e resulta de falha ou omissão.

A auditoria das demonstrações financeiras reconhece a possibilidade de ocorrência de fraude e por este motivo atribui ao profissional de auditoria a responsabilidade de se obter uma segurança razoável de que as demonstrações não contêm distorções relevantes, independente de erro ou fraude (CFC, 2016).

A NBC TA 200 (CFC, 2016) ressalta que, mesmo com a aplicação de procedimentos padronizados pelas normas de auditoria, não se chega a garantia absoluta de detecção de todas as distorções nas demonstrações financeiras. Isso se dá pelo fato das limitações inerentes ao processo de auditoria, em que se destacam a natureza do julgamento profissional do auditor e as características das evidências obtidas por ele.

Donald Cressey em sua teoria intitulada “triângulo da fraude” afirma que é necessário a presença simultânea de três elementos para a ocorrência da fraude (CRESSEY, 1953 apud ABDULLAHI e MANSOR, 2015).

Quadro 3 - Triângulo da fraude



Fonte: Adaptado de Abdullahi e Mansor (2015)

1. **Pressão** - refere-se a motivações pessoais que induzem ao comportamento fraudulento, e está frequentemente ligada a problemas financeiros.
2. **Oportunidade** - decorrente da existência de fragilidades, ou seja, brechas nos controles internos ou de uma governança falha, possibilitando ao fraudador cometer e ocultar a fraude.

- 3. Racionalização** - Envolve o processo psicológico, onde o indivíduo justifica o ato ilícito para si mesmo, formulando uma ideia moralmente aceitável utilizando de justificativas e desculpas para sustentar ação fraudulenta.

2.4. Mineração de dados na detecção de fraudes em demonstrações financeiras

De acordo com Kasztelnik e Jermakowicz (2024), a digitalização da contabilidade possibilita a aplicação de algoritmos para análise de dados que auxiliam na identificação de padrões anormais e irregularidades que acabam passando despercebidas pelos métodos de auditoria tradicionais.

A mineração de dados desempenha um importante papel neste contexto, aplicando análises preditivas e estatísticas direcionadas para identificação de riscos de fraudes e superando as limitações dos procedimentos de auditoria, especialmente com relação a capacidade de se analisar grandes bases de dados (KASZTELNIK e JERMAKOWICZ, 2024).

Métodos clássicos de auditoria utilizam de técnicas de amostragem quando necessário análise de grandes conjuntos de dados, e exige do profissional de auditoria um alto grau de julgamento profissional e senso crítico. Por outro lado, os modelos baseados em inteligência artificial são capazes de processar integralmente as essas bases de dados, aumentando consideravelmente as chances de detecção de fraudes complexas analisando todo o conjunto.

Apesar destes avanços, a interpretação dos resultados obtidos com técnicas de mineração exige cautela. Algoritmos podem apresentar vieses e falsos positivos, e por este motivo não se pode deixar de lado o julgamento profissional do auditor. O ideal é a combinação entre a tecnologia e habilidade profissional, de modo que se assegure a confiabilidade das decisões tomadas (KASZTELNIK e JERMAKOWICZ, 2024).

2.5. A lei de Newcomb-Benford

A lei de Newcomb-Benford, também chamada lei dos Números Anômalos, ou apenas lei de Benford descreve um fenômeno sobre a frequência de dígitos em conjuntos de dados numéricos. A lei descreve que, em coleções de valores naturais,

os dígitos iniciais não são distribuídos de forma uniforme como se pode pensar inicialmente.

Dados numéricos naturais são aqueles que ocorrem de forma espontânea, e representam razões de quantidades ou os tamanhos de fatos e eventos no mundo.

Este fenômeno foi observado inicialmente pelo astrônomo e matemático Simon Newcomb (1881) ao observar que, em tabelas de logaritmos impressas, as páginas referentes a números iniciados com o dígito 1 apresentavam um desgaste nitidamente maior do que as páginas finais como os números 7,8 e 9. A observação o levaram a concluir que dígitos menores apareciam com mais frequência no início dos números utilizados em cálculos científicos.

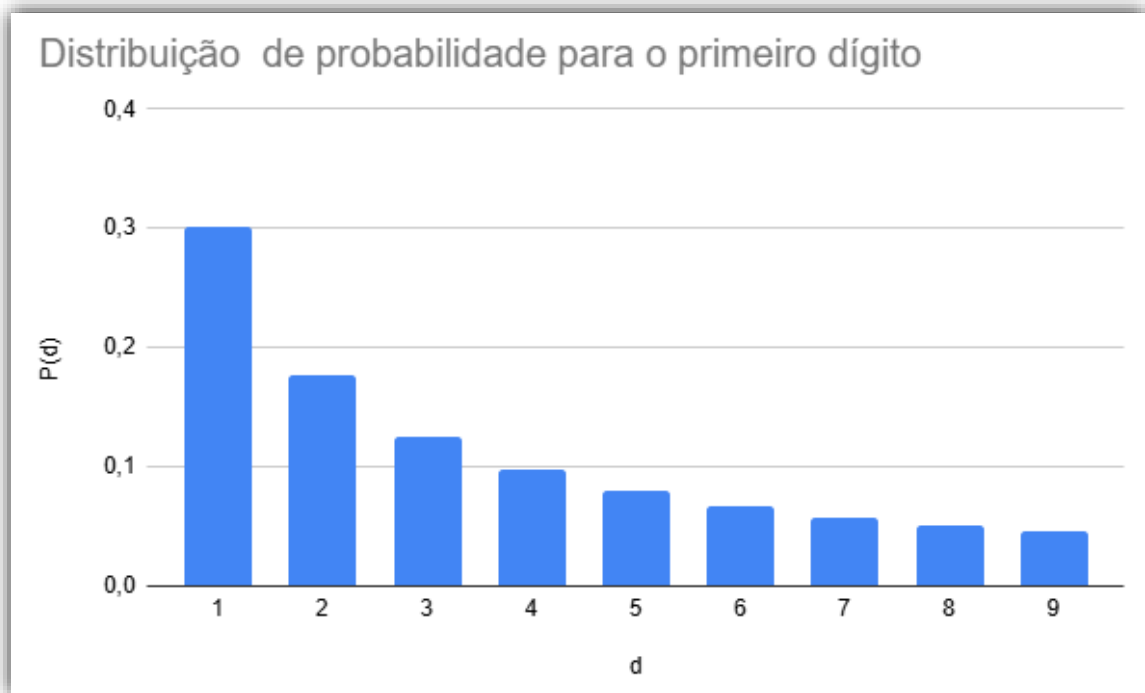
Mais tarde Benford (1938) redescobriu e expandiu o estudo de Newcomb, realizando testes de forma empírica, analisando 20.229 observações de vinte domínios distintos. Este estudo confirmou que a distribuição dos primeiros dígitos de fato seguia um padrão logarítmico previsível, em vez uma distribuição uniforme.

A lei de Benford estabelece que a probabilidade (P) de um dígito d_1 onde $d_1 \in \{1, 2, \dots, 9\}$, aparecer como o primeiro dígito significativo de um número é dada pela seguinte função logarítmica:

$$P(d_1) = \log\left(1 + \frac{1}{d_1}\right) \quad (1)$$

De acordo com essa função, o dígito 1 possui a maior probabilidade de ocorrência, aparecendo como primeiro dígito em aproximadamente 30,1% dos casos, diferentemente dos dígitos maiores que são menos frequentes à medida que se aproximam do dígito 9, sendo este aparecendo em apenas 4,6% das vezes.

Quadro 4 - Distribuição de probabilidade para o primeiro dígito



Fonte: Adaptado de Nigrini (2012)

A lei não se limita ao primeiro dígito, o padrão se estende de forma modificada, aos dígitos subsequentes. A probabilidade do segundo dígito d_2 é calculada por:

$$\text{Prob}(D_2 = d_2) = \sum_{d_1=1}^9 \log\left(1 + \frac{1}{d_1 d_2}\right); \quad d_2 \in \{0,1, \dots, 9\} \quad (2)$$

Conforme detalhado por Nigrini (2012), à medida que se avança para o terceiro e quartos dígitos, a distribuição se aproxima da uniformidade. O dígito 0, não é utilizado como primeiro dígito, mas torna-se o mais esperado como segundo dígito (certa de 11,97%). A Tabela 1 detalha as probabilidades esperadas para os quatro primeiros dígitos de um número natural.

Tabela 1 - Probabilidades esperadas dos dígitos de acordo com a lei de Benford

Dígito	1ª posição	2ª posição	3ª posição	4ª posição
0	—	0,11968	0,10178	0,10018
1	0,30103	0,11389	0,10138	0,10014
2	0,17609	0,10882	0,10097	0,10010
3	0,12494	0,10433	0,10057	0,10006
4	0,09691	0,10031	0,10018	0,10002
5	0,07918	0,09668	0,09979	0,09998
6	0,06695	0,09337	0,09940	0,09994
7	0,05799	0,09035	0,09902	0,09990
8	0,05115	0,08757	0,09864	0,09986
9	0,04576	0,08500	0,09827	0,09982

Fonte: Nigrini (2012)

A lei de Benford possui propriedades matemáticas fundamentais, sendo a “invariância de escala” a mais importante. Esta propriedade nos diz que a distribuição dos dígitos não muda, independente da unidade de medida adotada.

Por exemplo, um conjunto de dados financeiros que segue a lei de Benford em Reais (R\$) continuará a segui-la em outra unidade monetária (NIGRINI, 2012).

2.5.1. Aplicações em auditoria e detecção de fraudes

Enquanto conjunto de dados naturais tendem a seguir a lei de Benford, valores inventados ou criados artificialmente não seguem esse mesmo padrão. Ao criar números artificialmente para fraudar relatórios financeiros o ser humano tende a inventar de forma não aleatória, violando assim a distribuição esperada. Números inventados costumam ter excessos de certos dígitos como 7, 8 e 9, que podem parecer “mais aleatórios”, e uma escassez de outros, como o número 1.

Por isso a lei de Benford é uma importante ferramenta em auditoria, pois se espera que dados financeiros que reflitam transações reais, sigam a distribuição. Desvios exagerados desse padrão servem como um “alerta” (*red flag*), indicando áreas de risco que exigem investigação minuciosa (NIGRINI, 2012).

O método não identifica a fraude em si, mas aponta anomalias e distorções que podem ser resultado de erros, vieses sistêmicos ou de fato uma manipulação intencional.

Nigrini (2012) desenvolveu testes específicos que vão além do primeiro dígito, como a análise dos dois primeiros dígitos (*First-Two Digits*) e o teste de soma (*Summation Test*), para identificar números anormais, grandes ou duplicados que podem sinalizar fraude.

A lei de Benford exige características específicas dos dados, não sendo aplicável em todos os casos quando: conjuntos de dados com intervalos reduzidos, valores previamente delimitados ou influenciados por regras normativas podem não apresentar aderência ao padrão esperado. Por essas razões, sua aplicação deve ser utilizada como um indicador primário da existência de inconsistências, mas não como prova conclusiva e incontestável de fraude (ORTH, MICHAELSEN e LERNER, 2020).

2.5.2. Testes Estatísticos de Conformidade

Para comprovação se um conjunto de dados segue ou não a distribuição prevista pela lei de Benford, foram desenvolvidos testes estatísticos de aderência. Esses métodos permitem comparar as frequências observadas dos dígitos com as frequências esperadas, fornecendo uma medida quantitativa de conformidade. Entre os mais utilizados:

Teste Z (Z-score): avalia a discrepância entre a frequência observada e a frequência esperada de cada dígito de forma individual. O resultado é expresso em forma de desvios-padrão, sendo que elevados valores absolutos de Z indicam maior probabilidade de que o dado analisado não siga a distribuição esperada. Esse teste é especialmente útil para identificar anomalias pontuais em dígitos específicos.

Teste do Qui-quadrado (χ^2): examina a aderência global da amostra com relação a distribuição de Benford. Ele considera todos os dígitos simultaneamente, somando os desvios quadráticos ponderados entre frequências observadas e esperadas. Por ser robusto tem sido amplamente utilizado em conjuntos de dados volumosos, permitindo avaliar a consistência geral da distribuição (NIGRINI, 2012).

Teste da Soma: desenvolvido por Nigrini, consiste em somar as diferenças entre as frequências observadas e esperadas dos dígitos, considerando o sinal positivo ou negativo de cada desvio. Esse teste é mais sensível a tendências sistemáticas de manipulação, já que desvios em uma direção específica como por exemplo, excesso de dígitos baixos ou altos, não se anulam.

A utilização combinada destes métodos aumenta a confiabilidade das conclusões. Enquanto o teste Z permite localizar anomalias específicas, o qui-quadrado fornece uma visão global da aderência, e o teste da soma contribui para detectar padrões sistemáticos de distorção. Dessa forma, a integração dos testes possibilita um diagnóstico mais abrangente, com equilíbrio entre precisão local, análise global e detecção de vieses ocultos nos dados.

Com o embasamento teórico adquirido lei de Benford e os testes estatísticos, o próximo passo é definir como aplicar esses conceitos de forma prática. O próximo capítulo descreve a metodologia adotada neste trabalho, detalhando as etapas de coleta, tratamento e análise dos dados referentes aos repasses extraorçamentários do Estado de Goiás.

3. METODOLOGIA

Fraudes nas demonstrações financeiras e em gastos públicos representam um desafio significativo para a governança e a transparência no setor público brasileiro, comprometendo a alocação eficiente de recursos e a confiança da sociedade nas instituições. Gastos do governo com despesas extraorçamentárias como, transferências não previstas no orçamento inicial e que são autorizadas por motivos emergenciais ou excepcionais são necessários para atender de forma rápida a demandas sociais. Essa flexibilidade na utilização de recursos não previsto alinhado aos elementos do triângulo da fraude (pressão, oportunidade e racionalização) torna as contas vulneráveis a manipulação fraudulenta.

Diante deste cenário, este trabalho pretende abordar como a aplicação da lei de Benford e das técnicas de mineração de dados podem auxiliar nos procedimentos de auditoria das demonstrações financeiras, na identificação de distorções e indícios de fraude. Com objetivo de analisar se um conjunto de dados financeiros atende as distribuições de dígitos observados pela lei de Benford e possibilitar ao profissional de auditoria chegar a uma conclusão se o conhecimento adquirido pode ser utilizado para embasar decisões relacionadas a presença de fraude.

Neste trabalho adotou-se uma abordagem quantitativa/exploratória, aplicada em um estudo de caso, alinhada com os princípios da ciência de dados. O estudo de caso concentra-se na análise de um conjunto de dados referente a transferências extraordinárias de valores do Estado de Goiás em 2024, utilizando um conjunto de dados públicos para validar sua conformidade com a lei de Benford.

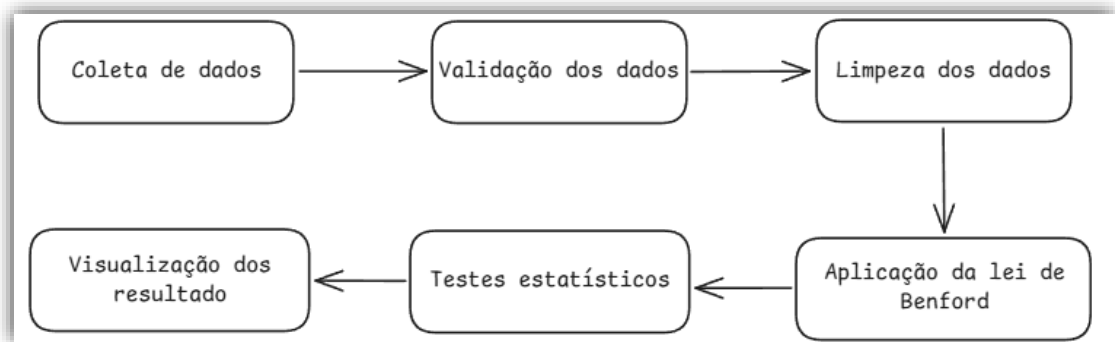
Conforme ilustra o Quadro 5, o desenvolvimento deste trabalho seguiu o fluxo baseado nas etapas do processo KDD:

- **Coleta:** Os dados referentes aos repasses extraorçamentários foram obtidos diretamente no Portal da Transparência do Estado de Goiás;
- **Validação e limpeza:** Utilizou-se a linguagem Python para validação inicial, padronização de valores monetários e remoção de registros inconsistentes que não se aplicam a lei de Benford;
- **Aplicação da lei de Benford:** A etapa de mineração dos dados, onde foi implementado algoritmos para extração dos primeiros dígitos e cálculo das

frequências observadas, comparando com a distribuição previstas pela lei de Benford;

- **Testes estatísticos e visualização:** Por fim seguiu-se com o pós-processamento, onde os resultados da mineração foram submetidos a testes estatísticos e gerado representações gráficas para interpretação e validação das anomalias encontradas.

Quadro 5 - Etapas adotadas no estudo de caso

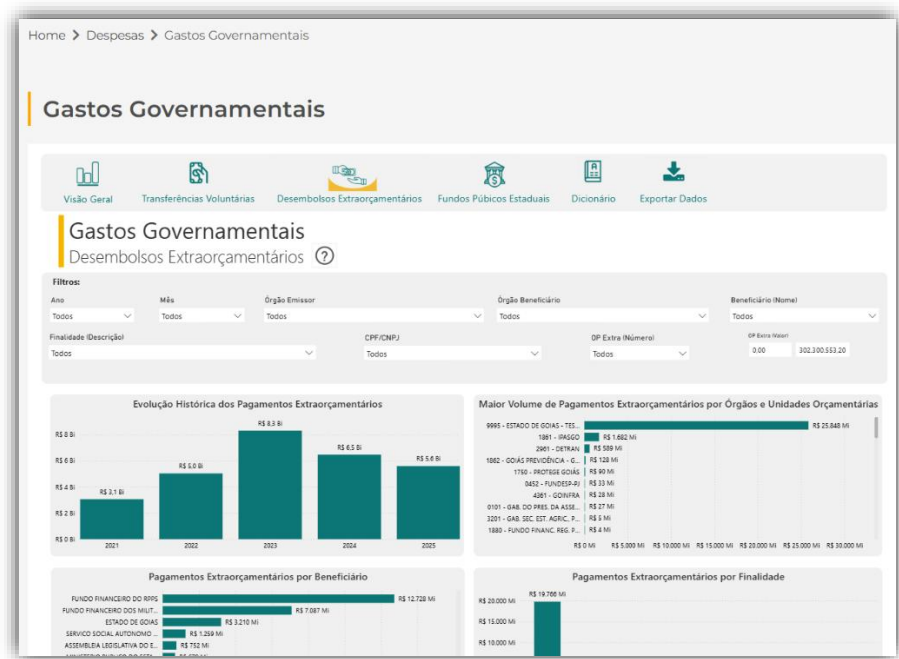


Fonte: O autor

3.1. Coleta de Dados

O conjunto de dados utilizado foi obtido do **Portal da Transparência do Estado de Goiás**, que disponibiliza informações públicas sobre receitas, despesas e transferências financeiras do governo estadual, disponíveis na seção **Gastos Governamentais – Desembolsos Extraorçamentários**, conforme Quadro 6 retirado do site. Desembolsos Extraorçamentários compreende os pagamentos que não estavam previstos no processo de execução orçamentária.

Quadro 6 - Portal da transparência - Gastos Governamentais



Fonte: Portal da transparência do estado de Goiás

Os dados foram baixados em formato CSV da página oficial (<https://transparencia.go.gov.br/gastos-governamentais/>) em 25/09/2025. Convertido para .xlsx utilizando o Excel e simplificado nome das colunas para melhor visualização e identificação dos dados relevantes, conforme mostrado no Quadro 7.

Na coleta garantiu-se que os dados mantivessem sua integridade original, sem manipulação prévia, apenas armazenados para posterior validação e manipulação conforme nas demais etapas do processo.

Quadro 7 - Dados extraídos e organizados

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Exercício	Mês	Finalidade	Finalidade (Descrição)	Beneficiário	CPF	OP Extra	Órgão Emissor	Órg	Valor	Data Atual	CPF	Desc
2	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4779	9995 - ESTADO	0	878,44	#####	02.29	Agosto
3	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4793	9995 - ESTADO	0	1362,81	#####	02.29	Agosto
4	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4794	9995 - ESTADO	0	931,51	#####	02.29	Agosto
5	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4807	9995 - ESTADO	0	505,42	#####	02.29	Agosto
6	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4822	9995 - ESTADO	0	785,72	#####	02.29	Agosto
7	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4823	9995 - ESTADO	0	1942,86	#####	02.29	Agosto
8	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4829	9995 - ESTADO	0	1510,67	#####	02.29	Agosto
9	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4839	9995 - ESTADO	0	622	#####	02.29	Agosto
10	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4852	9995 - ESTADO	0	1395,17	#####	02.29	Agosto
11	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4857	9995 - ESTADO	0	505,42	#####	02.29	Agosto
12	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4872	9995 - ESTADO	0	2428,2	#####	02.29	Agosto
13	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4873	9995 - ESTADO	0	1128,15	#####	02.29	Agosto
14	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4875	9995 - ESTADO	0	616,07	#####	02.29	Agosto
15	2024	8	74	Restituição de Depósito	TRIBUNAL DE JUSTICA	####	4914	9995 - ESTADO	0	4281,91	#####	02.29	Agosto

3.2. Apresentação e validação dos dados

Esta é uma etapa essencial para entender como está organizado os atributos identificar os ajustes necessários à adequação dos dados para aplicação dos princípios de Benford. Realizou-se utilizando a linguagem *Python*, e o ambiente integrado *Jupyter Notebook*, incluso na plataforma *Anaconda*.

A escolha da ferramenta se deu por sua ampla utilização em pesquisas científicas e pelo conjunto completo de ferramentas voltadas para a ciência de dados, facilidade de instalação, e de ter as ferramentas necessárias para o trabalho em um único local.

A plataforma Anaconda é voltada para análise de dados e aprendizado de máquina, e contém uma série de bibliotecas especializadas que facilitam a manipulação e o tratamento de dados.

O *Jupyter Notebook*, oferece um ambiente de codificação *Python* que permite a execução de código, visualização de resultados e conta com um sistema de arquivos integrado, oferecendo praticidade na para execução de código.

As principais bibliotecas *Python* utilizadas nesta etapa foram:

- **Pandas** - para manipulação e análise de dados em tabelas;
- **NumPy** - para realizar operações matemáticas;
- **Pathlib** - para gerenciamento do sistema de arquivos.

Inicialmente realizou-se a conversão de formato .csv para .xlsx possibilitando seleção manual das colunas relevantes do conjunto original, mantendo apenas as informações pertinentes ao objetivo do estudo. A exemplo foram removidas colunas de identificação, como CNPJ, CNPJ mascarado e CPF.

O documento passou a conter apenas atributos (colunas) suficientes para identificação da movimentação financeira e seu valor. As colunas mantidas e renomeadas para melhor identificação, foram:

- **beneficiário**: identificação da entidade ou órgão que recebeu o repasse;
- **numero_operação**: número identificador da operação de repasse;
- **valor_repasse**: valor monetário do repasse, em reais. Sendo este o dado principal no qual foi trabalhos nas etapas seguintes.

Tabela 2 - Amostra dos cinco primeiros registros do arquivo após ajustes

beneficiário	numero_operação	valor_repasse
Tribunal de Justiça do Estado de Goiás	4779	878,44
Tribunal de Justiça do Estado de Goiás	4793	1362,81
Tribunal de Justiça do Estado de Goiás	4794	931,51
Tribunal de Justiça do Estado de Goiás	4807	505,42
Tribunal de Justiça do Estado de Goiás	4822	785,72

Fonte: O autor

A Tabela 2 mostra como cada transação está organizada dentro do conjunto de dados. As colunas disponíveis permitem que, caso uma anomalia seja detectada pela lei de Benford, o auditor consiga ter um grau de rastreabilidade, chegando à um conjunto de transações para averiguação.

Após a organização dos dados, foi executado o código em *Python*, responsável por padronizar os valores monetários e retornar informações sobre os dados, necessárias para validar a conformidade do conjunto com os requisitos necessários para aplicação da lei de Benford. (Todos os arquivos de código fonte utilizados neste trabalho estão disponíveis em repositório *github* que podem ser acessado em <https://github.com/jpsoul-dev/codebase>)

Também foi realizada a padronização dos valores para o padrão americano, utilizando “.” como separador decimal para facilitar a manipulação pelas bibliotecas, visto que utilizam este padrão.

Como resultado da execução foi possível extrair as seguintes informações:

Valores negativos: 1

Valores iguais a zero: 3.063

Valores não numéricos (incluindo em branco): 2

Ordens de magnitude distintas: 12

Quantidade de registros: 43786

Na análise foi dispensada a identificação de *outliers* pois não interferem negativamente na aplicação de Benford. Por outro lado, essa validação inicial mostra

que o conjunto possui uma boa variedade de ordens de magnitude estando adequado neste quesito.

Os dados são de transações financeiras reais, desta forma já se espera que os valores sejam números naturais, diferentemente de conjuntos de dados gerados artificialmente que simulam transações e não são adequados para aplicação de Benford.

Com essas informações conclui-se que o conjunto de dados original, está apenas parcialmente adequado à aplicação da lei de Benford, sendo necessário a realização de limpeza para remoção de valores zerados e não numéricos.

Com relação aos valores negativos optou-se por mantê-los convertendo em positivo (módulo). Esta escolha se justifica devido aos lançamentos contábeis serem diferenciados em função da natureza da operação (débito ou crédito).

3.3. Limpeza dos Dados

Nesta etapa, realizou-se a limpeza do conjunto de dados, utilizando a biblioteca *Pandas*, para garantir que os dados estejam adequados à aplicação da lei de Benford na etapa de mineração. O processo de limpeza foi organizado em quatro rotinas:

1. **Remoção de valores vazios:** Foram identificados e excluídos os registros onde o campo `valor_repasse` estava em branco;
2. **Filtragem de valores não numéricos:** Foram removidos registros contendo caracteres textuais no campo `valor_repasse`, como demonstrado na Tabela 3;
3. **Exclusão de valores zerados:** Registros com valor igual a 0,00 foram removidos, visto que o zero não possui dígito significativo, não sendo aplicável portanto à análise da lei de Benford;
4. **Padronização de valores negativos:** Foi identificado a presença de registros negativos. Como a lei de Benford analisa a frequência dos dígitos independente do sinal, aplicou-se uma função de módulo para converter esses registros em positivos, preservando a informação para o cálculo conforme exemplificado na Tabela 4 para o único valor negativo encontrado.

Ao final do processamento, obteve-se o seguinte resultado na redução do conjunto de dados:

Valores zerados removidos: 3063
 Valores em branco: 1
 Valores não numéricos: 1
 Valores convertidos em módulo: 1
 Total de registros removidos: 3065
 Total mantido: 40720
 Percentual mantido: 93,0%

Tabela 3- exemplo de registros removidos

beneficiário	numero_operação	valor_repasse
Tribunal de Justiça do Estado de Goiás	6237	0,00
Secretaria de Estado da Segurança Pública	1300	
Secretaria de Estado da Segurança Pública	2074	“adfsd”

Fonte: O autor

Tabela 4 - Registro negativos que foi convertido em positivo

beneficiário	numero_operação	valor_repasse
Secretaria de Estado da Segurança Pública	1695	- 3.705,38

Fonte: O autor

A remoção de valores zerados e a conversão de negativos, são necessários para a função logarítmica de Benford. Esses registros poderiam inflar algumas frequências de forma errônea ou gerar erros de cálculo (como log de zero por exemplo), invalidando os testes de hipótese propostos. Isso mostra a necessidade de preparar o conjunto de dados para evitar distorções no modelo matemático previsto.

Após o processo de limpeza e realização de uma nova validação sobre os dados. O resultado mostrou-se adequado, atendendo às premissas da lei de Benford e as exigências necessárias para os testes estatísticos de conformidade.

3.4. Aplicação da lei de Benford

Com base no conjunto de dados tratado anteriormente, iniciou-se a etapa de mineração de dados propriamente dita, aplicando a lei de Benford. A etapa computacional primária consistiu na extração do primeiro dígito significativo (D1) e do segundo dígito (D2) de cada transação da coluna **valor_repasse**.

Para isso foi criado um código *Python*, que implementa a lógica de extração. O processo aplicado a cada valor do conjunto é descrito como:

1. O valor é convertido em *string* e o separador decimal "." é removido.
2. Suprimiu os zeros à esquerda da *string* resultante. Esta etapa é crucial para garantir que os dígitos sejam corretamente identificados em valores decimais inferiores a 1.00.
3. O primeiro caractere da *string* tratada é capturado como D1 e o segundo, caso exista, como D2.

O Quadro 8 mostra o resultado da extração do primeiro e segundo dígito é exemplo do arquivo final preparado.

Quadro 8 - Resultado da extração do primeiro e segundo dígito

```
Extração de dígitos concluída!
Primeiro dígito extraído:      40,720 registros
Segundo dígito extraído:      40,706 registros
Registros com apenas 1 dígito: 14

Arquivo gerado: dados_limpos_benford.xlsx
=====

Primeiros 10 registros do arquivo final:
```

beneficiario	numero_operacao	valor_repasse	primeiro_digito	segundo_digito
TRIBUNAL DE JUSTICA DO ESTADO DE GOIAS	4779	878.44	8	7.0
TRIBUNAL DE JUSTICA DO ESTADO DE GOIAS	4793	1362.81	1	3.0
TRIBUNAL DE JUSTICA DO ESTADO DE GOIAS	4794	931.51	9	3.0
TRIBUNAL DE JUSTICA DO ESTADO DE GOIAS	4807	505.42	5	0.0
TRIBUNAL DE JUSTICA DO ESTADO DE GOIAS	4822	785.72	7	8.0
TRIBUNAL DE JUSTICA DO ESTADO DE GOIAS	4823	1942.86	1	9.0
TRIBUNAL DE JUSTICA DO ESTADO DE GOIAS	4829	1510.67	1	5.0
TRIBUNAL DE JUSTICA DO ESTADO DE GOIAS	4839	622.00	6	2.0
TRIBUNAL DE JUSTICA DO ESTADO DE GOIAS	4852	1395.17	1	3.0
TRIBUNAL DE JUSTICA DO ESTADO DE GOIAS	4857	505.42	5	0.0

Fonte: O autor

Com os dados preparados, procedeu-se à análise de conformidade de D1. O método compreendeu as seguintes etapas:

1. cálculo das frequências observadas dos dígitos iniciais (1–9);

2. comparação entre frequência observada e a esperada segundo Benford;
3. geração de representações gráficas, utilizadas para análise visual inicial.

As principais novas bibliotecas adicionais utilizadas nesta etapa foram:

- **matplotlib** e **seaborn** - para gerar de gráficos e elementos visuais;
- **scipy.stats** - para funções estatísticas X^2 e probabilidades;
- **pathlib** / **openpyxl** - controle de arquivos e manipulação de planilhas Excel.

Pela Tabela 5 pode-se comparar a distribuição observada e a esperada pela lei de Benford, juntamente com a diferença absoluta.

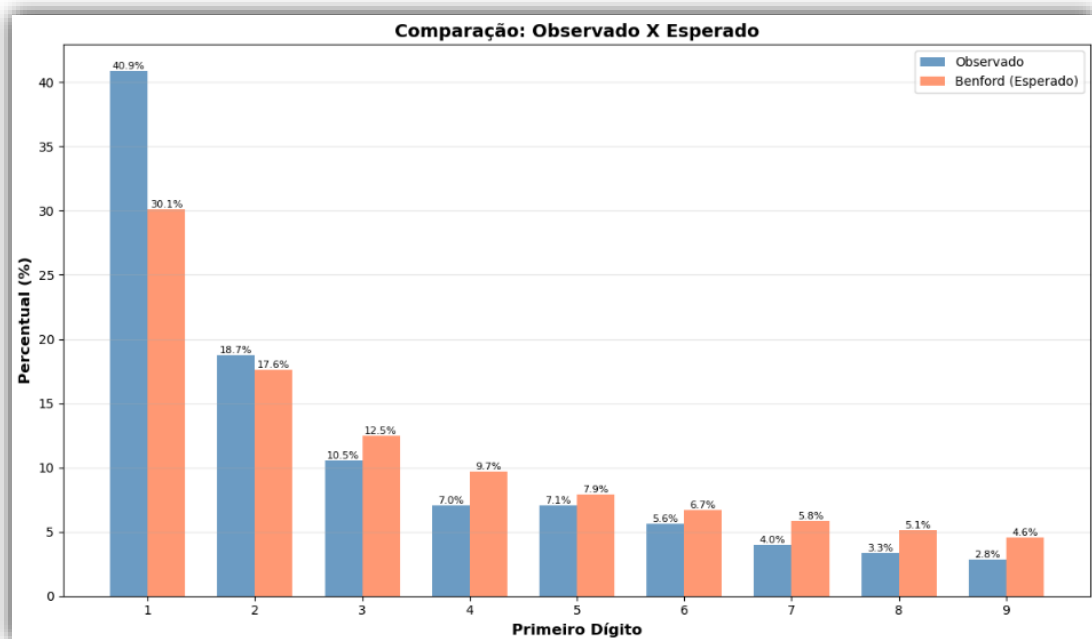
Tabela 5 - Distribuição observada e esperada do primeiro dígito

Dígito	Frequência observada	Percentual observado (%)	Frequência esperada	Percentual esperado (%)	Diferença (p.p.)
1	16.648	40,88	12.258	30,10	+10,78
2	7.622	18,72	7.170	17,61	+1,11
3	4.287	10,53	5.088	12,49	-1,97
4	2.869	7,05	3.946	9,69	-2,65
5	2.872	7,05	3.224	7,92	-0,87
6	2.274	5,58	2.726	6,69	-1,11
7	1.630	4,00	2.361	5,80	-1,80
8	1.363	3,35	2.083	5,12	-1,77
9	1.155	2,84	1.863	4,58	-1,74

Fonte: O autor

O Quadro 9 demonstra de forma gráfica a comparação entre percentual esperado e observado.

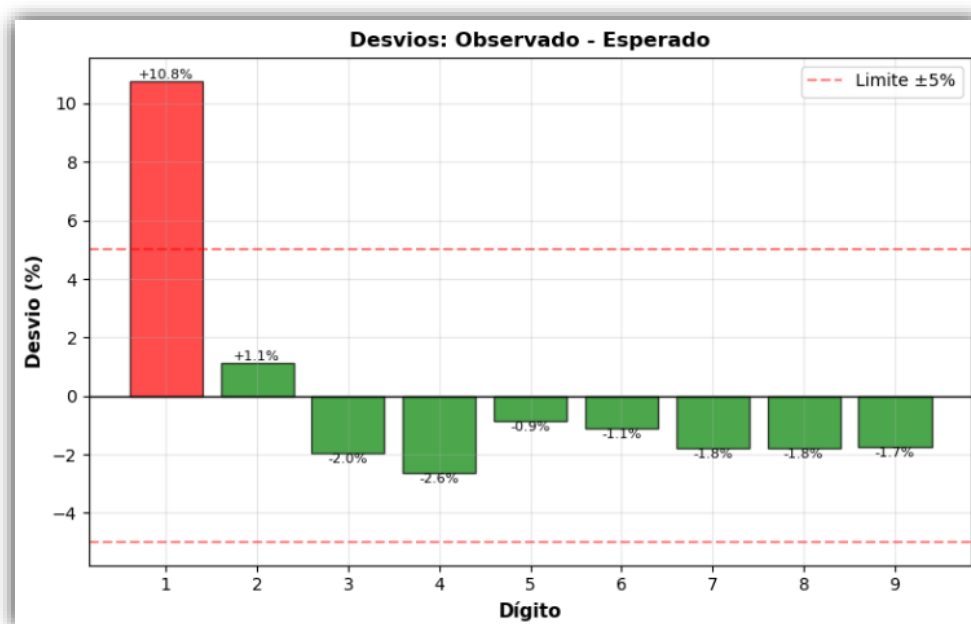
Quadro 9 - Comparação entre distribuição observada e esperada



Fonte: O autor

O Quadro 10 apresenta os desvios entre a frequência observada e a esperada. Para fins de análise preliminar de auditoria, estabeleceu-se um intervalo de tolerância de $\pm 5\%$, demonstrado no gráfico pelas linhas tracejadas em vermelho, as barras que ultrapassam este limite indicam anomalias.

Quadro 10 - Gráficos dos desvios



Fonte: O autor

Para confirmação de que os desvios apresentados são estatisticamente significantes ou fruto do acaso, aplicam-se os testes de hipótese descritos na seção a seguir.

3.5. Testes estatísticos

Esta etapa apresenta os testes estatísticos formais para quantificar a significância das anomalias observadas inicialmente e a formulação das hipóteses.

Todos os testes foram feitos utilizando *Python* e o código fonte está disponível no repositório *github* <https://github.com/jpsoul-dev/codebase>.

Hipótese Nula (H0): A distribuição dos primeiros dígitos das transações financeiras segue a distribuição da lei de Benford.

Hipótese Alternativa (H1): A distribuição dos primeiros dígitos não segue a distribuição da lei de Benford.

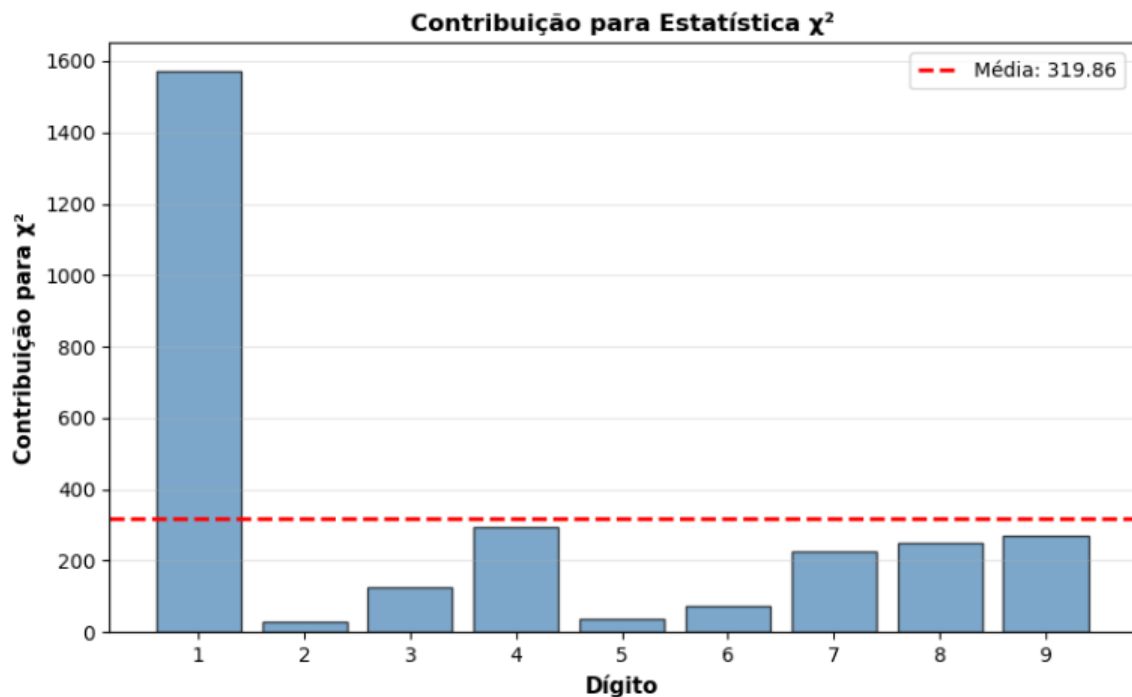
3.5.1. Teste de aderência Qui-Quadrado (χ^2)

O teste qui-quadrado (χ^2) é o principal teste estatístico para avaliar a conformidade geral de todo o conjunto de dados com a lei de Benford. Este teste compara a frequência de dados observada em cada categoria (dígitos 1-9) com a frequência esperada e demonstrada anteriormente.

Para o teste χ^2 o nível de significância adotado foi de 0,05 (5%), com 8 graus de liberdade ($GL = k - 1$, onde $k = 9$ dígitos). O valor crítico conforme tabela de distribuição χ^2 para estas condições é de 15,51.

Ao calcular a estatística do teste, obteve-se um valor χ^2 total elevado. O Quadro 11 detalha a contribuição individual de cada dígito para este somatório.

Quadro 11 - Contribuição de cada dígito para a estatística



Fonte: O autor

3.5.2. Teste Z-score

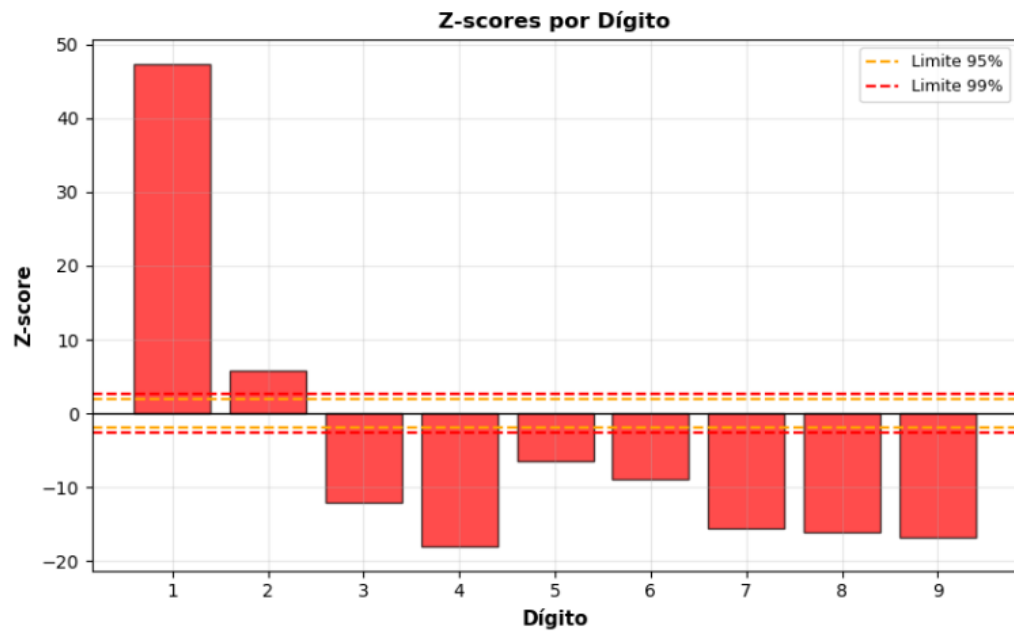
Enquanto o Qui-Quadrado avalia o conjunto como um todo, o Teste Z-score foi aplicado para testar a cada dígito de forma individual. O Z-score mede quantos desvios padrão a frequência observada se afasta da esperada. Temos que:

- Para um nível de confiança de 95%, o valor crítico é $|Z| > 1,96$.
- Para um nível de confiança de 99%, o valor crítico é $|Z| > 2,57$.

O Quadro 12 mostra o resultado dos Z-scores calculados para cada dígito.

No gráfico, as linhas tracejadas vermelha e laranja, são os limites crítico de $\pm 1,96$ e $\pm 2,57$ e inicialmente pode-se observar que todos os dígitos extrapolaram estes limites.

Quadro 12 - Gráfico de z-scores por dígito



Fonte: O autor

4. RESULTADOS

A proposta metodológica utilizando o processo de KDD e aplicação da lei de Benford sobre os dados de desembolsos extraorçamentários do Estado de Goiás permitiu obter, processar e analisar 40.720 registros de repasses válidos.

Utilizou-se dos resultados obtidos nas etapas anteriores para comparação entre as frequências observadas e a esperada conforme dita a lei de Benford, validada pelos testes de hipótese estatísticos.

4.1. Análise da distribuição de frequências

Conforme dados apresentados na Tabela 5 e a comparação gráfica entre frequência esperada e observada no Quadro 9, a primeira evidência de não conformidade é a distribuição dos primeiros dígitos, que se afastam do comportamento logarítmico esperado.

O destaque mais evidente é para o dígito 1, que, pela lei de Benford se esperava em torno de 30,10% dos registros iniciando com 1. Entretanto a frequência observada foi de 40,88%, o que resulta em um desvio positivo de 10,78 pontos percentuais.

Mapeando para a quantidade representativa de registros, temos:

Quantidade Esperada (E_1): $E_1 = 40.720 \times 0,30103 \approx 12.257,94$

Quantidade Observada (O_1): $O_1 = 40.720 \times 0,4088 \approx 16.646,33$

O que representa em termos absolutos uma diferença de mais de 4.000 registros além do previsto.

O gráfico dos desvios percentuais ilustrado no Quadro 10, revela melhor a magnitude das discrepâncias em relação ao modelo teórico. Reforçando o comportamento anormal observado para o dígito 1, extrapolando o limite de tolerância estabelecido ($\pm 5\%$) sugerindo uma concentração artificial de valores nessa faixa inicial.

Em contrapartida observa-se o inverso para os dígitos de 3 a 9, apresentando desvios negativos

Uma rápida análise pelo somatório dos desvios negativos e positivos revela um forte indicativo de efeito compensatório matematicamente preciso, ao excesso do dígito 1.

Desvios negativos (dígitos de 3 a 9): -11,9%

Desvios positivos (dígitos 1 e 2): 11,9%

4.2. Análise do teste Qui-Quadrado

O resultado do teste χ^2 demonstrado no Quadro 11, mostra que o valor calculado para a estatística foi muito superior aos limites de referência, o que leva a rejeição da Hipótese Nula (H_0).

A decomposição do teste permitiu identificar como principal fonte da divergência, o dígito 1, sendo responsável pela maior parcela do valor total do teste, contribuindo com mais de 1.500 pontos.

Desta forma pode-se afirmar que a falta de adequação do modelo não é distribuída de forma uniforme entre os dígitos, mas sim concentrada na distorção dos valores iniciados em 1.

4.3. Análise do teste Z-score

O Teste Z-score mostra a conformidade de cada dígito de forma individual, mostrado no Quadro 12. Na análise do gráfico gerado destaca-se que o dígito 1 obteve um Z-score de 47,0, este valor excede muito além do intervalo de confiança ($\pm 2,57$).

Desta forma pode-se afirmar estatisticamente que a frequência desse dígito não é compatível com a distribuição esperada, e não sendo passível de atribuição ao acaso.

Os demais dígitos também apresentam Z-scores fora do intervalo de confiança, sendo que os dígitos de 3 a 9 apresentam valores negativos inferiores a -2,57, o que confirma a baixa utilização desses números, indicando possível manipulação sistemática. Mesmo os dígitos 2 e 6 com desvios menores ainda se situam na zona de rejeição da hipótese nula. Assim, pode-se concluir que a não conformidade afeta toda a distribuição.

4.4. Síntese dos resultados

Os testes estatísticos aplicados fornecem evidências de que os valores de repasses extraorçamentários analisados não atendem a lei de Benford.

O resultado obtido nos testes Z-score e Qui-quadrado mostram a eficácia destas ferramentas computacionais na detecção de desvios padrões.

A metodologia aplicada foi capaz de identificar em um conjunto de mais de 40 mil transações que existe um risco predominante associado aos valores iniciados com dígito 1. Isso comprova a utilidade da lei de Benford na auditoria contábil, como um teste preliminar capaz de separar dados anômalos que exigem uma verificação detalhada, de dados que seguem o comportamento esperado, dando suporte a análise crítica e julgamento profissional em procedimentos de auditoria.

As evidências estatísticas e as distorções identificadas fornecem a base necessária para responder aos objetivos propostos neste trabalho. O próximo capítulo apresenta as conclusões e as recomendações para trabalhos futuros no contexto de aplicação da ciência de dados na auditoria das demonstrações financeiras.

5. CONCLUSÃO

A proposta deste trabalho teve como objetivo validar a aplicação da lei de Benford como uma ferramenta auxiliar em auditorias contábeis, utilizando como caso de uso a aplicação sobre os repasses extraorçamentários do Estado de Goiás no exercício de 2024. Partindo da premissa de que os dados de transações devem seguir a distribuição logarítmica natural dos dígitos iniciais, quando estão isentos de manipulações.

Utilizando de um processo baseado em KDD, foi possível a partir das análises estatísticas fundamentais qui-quadrado e z-score demonstrar divergências críticas entre frequência observada e esperada, com destaque para o excesso do dígito 1 como demonstrado na sessão de análise de resultados, rejeitando assim a hipótese nula de conformidade.

Desta forma concluiu-se que o conjunto de dados analisado apresenta distorções significativas. Contudo é importante ressaltar, que a não conformidade com a lei de Benford aponta para a existência de anomalias estatísticas, mas não deve ser utilizada por si só, como prova definitiva de fraude. A ferramenta atua como um sinalizador de riscos (uma *red flag*), orientando onde a auditoria deve focar maior parte de seus esforços, melhorando o processo e proporcionando ao auditor sair de uma análise aleatória para um foco direcionado nos registros que violam o padrão matemático esperado, otimizando o tempo e recursos gastos nos procedimentos de auditoria.

Referente as distorções constatadas, levantam-se algumas hipóteses causais principais que podem explicar os desvios identificados nos dígitos:

1. Fragmentação de repasses: a divisão de pagamentos em valores menores, o que pode gerar uma concentração artificial de determinados dígitos;
2. Contratos pré-fixados: a existência de acordos entre órgãos e o Tesouro Estadual com valores exatos e recorrentes, distorcendo a distribuição natural esperada;
3. Manipulação intencional: a possibilidade real de fraude ou adulteração de registros com intuito de mascarar desvios de recursos.

Considerando que na preparação inicial dos dados foi realizada também a extração do segundo dígito, um aprofundamento futuro esperado deste trabalho seria naturalmente a realização de testes para o segundo dígito e de dígitos combinados.

Embora não tenha sido explorado no escopo deste trabalho, a aplicação de testes para os dois primeiros dígitos e análise isolada do segundo dígito, podem revelar padrões de manipulação diferentes e mais sutis, não demonstrados com a análise apenas do primeiro dígito. Seguindo como premissa de que os demais números também podem ter sido manipulados, principalmente em valores maiores.

Outra sugestão para estudo futuro, considerando a utilização do mesmo conjunto de dados, é a segmentação por beneficiário do repasse, afim de identificar se as anomalias se concentram em órgãos específicos, especialmente importante na auditoria para rastreabilidade de quem contribui para as fraudes.

Também se propõe o uso de dados rotulados pela auditoria humana para treinar modelos supervisionados de Aprendizado de Máquina. Dessa forma, o sistema evoluiria de apenas análise estatística para preditiva, capaz de reconhecer padrões complexos de fraude, aumentando significativamente a precisão na detecção de manipulação nas demonstrações financeiras.

REFERÊNCIAS

ABDULLAHI, R.; MANSOR, N. Fraud Triangle Theory and Fraud Diamond Theory: Understanding the Convergent and Divergent for Future Research. **International Journal of Academic Research in Accounting, Finance and Management Sciences**, v. 5, n. 4, 2015.

BENFORD, F. The law of anomalous numbers. **Proceedings of the American Philosophical Society**, v. 78, n. 4, 1938.

CONSELHO FEDERAL DE CONTABILIDADE. **NBC TA 200**: objetivos gerais do auditor independente e a condução da auditoria em conformidade com normas de auditoria. Brasília, DF: CFC, 2009.

CONSELHO FEDERAL DE CONTABILIDADE. **NBC TA 240**: responsabilidade do auditor em relação a fraude, no contexto da auditoria de demonstrações contábeis. Brasília, DF: CFC, 2009.

CONWAY, Drew. **The Data Science Venn Diagram**. 2013. Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>. Acesso em: 05 nov. 2025.

CORREA, Ana Paula Araujo; QUEIROZ, Eder; TREVISAN, Newton. **Teste do qui-quadrado**. Curitiba: Universidade Federal do Paraná, [s.d.]. Material didático da disciplina CE001 – Bioestatística.

GOIÁS (Estado). Secretaria da Economia. **Portal da Transparência do Estado de Goiás**: gastos governamentais – desembolsos extraorçamentários. [2024]. Disponível em: <https://transparencia.go.gov.br/gastos-governamentais/>. Acesso em: 25 set. 2025.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining**: um guia prático. 1. ed. Rio de Janeiro: Elsevier, 2005.

GRUS, Joel. **Data Science do zero**: primeiras regras com o Python. Tradução: Welington Nascimento. Rio de Janeiro: Alta Books, 2016.

KASZTELNIK, Karina; JERMAKOWICZ, E. K. Financial Statement Fraud Detection in the Digital Age: Artificial Intelligence versus Traditional Methodologies. **The CPA Journal**, jun. 2024. Disponível em: <https://www.researchgate.net/publication/381460085>. Acesso em: 09 set. 2025.

NEWCOMB, S. Note on the frequency of use of the different digits in natural numbers. **American Journal of Mathematics**, v. 4, n. 1, 1881.

NIGRINI, M. J. **Benford's Law**: applications for forensic accounting, auditing, and fraud detection. Hoboken: John Wiley & Sons, 2012.

ORTH, C. de O.; MICHAELSEN, A. T.; LERNER, A. F. Lei de Newcomb-Benford e auditoria contábil: uma revisão sistemática de literatura. **Revista Gestão e Desenvolvimento**, v. 17, n. 2, 2020.

PROVOST, Foster; FAWCETT, Tom. Data Science and Its Relationship to Big Data and Data-Driven Decision Making. **Big Data**, v. 1, n. 1, mar. 2013.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R**. Rio de Janeiro: Elsevier, 2016.

RESOLUÇÃO nº 038/2020 — CEPE

ANEXO I

APÊNDICE ao TCC

Termo de autorização de publicação de produção acadêmica

O(A) estudante João Paulo Teodoro Pereira do Curso de Engenharia de computação, matrícula 20121003303983, telefone: 62 98641 5387 e-mail ipsoul.dev@gmail.com, na qualidade de titular dos direitos autorais, em consonância com a Lei nº 9.610/98 (Lei dos Direitos do autor), autoriza a Pontifícia Universidade Católica de Goiás (PUC Goiás) a disponibilizar o Trabalho de Conclusão de Curso intitulado Estudo da Ciência de Dados na detecção de fraude em demonstrações financeiras, gratuitamente, sem ressarcimento dos direitos autorais, por 5 (cinco) anos, conforme permissões do documento, em meio eletrônico, na rede mundial de computadores, no formato especificado (Texto (PDF); Imagem (GIF ou JPEG); Som (WAVE, MPEG, AIFF, SND); Vídeo (MPEG, MOV, AVI, QT); outros, específicos da área; para fins de leitura e/ou impressão pela internet, a título de divulgação da produção científica gerada nos cursos de graduação da PUC Goiás.

Goiânia, 30 de Setembro de 2025

Assinatura do(s) autor(es): _____
 Documento assinado digitalmente
JOÃO PAULO TEODORO PEREIRA
Data: 30/09/2025 16:12:54-0300
Verifique em <https://validar.it.gov.br>

Nome completo do autor: João Paulo Teodoro Pereira

Assinatura do professor-orientador: _____
 Documento assinado digitalmente
ANDRÉ LUIZ ALVES
Data: 30/09/2025 20:10:11-0300
Verifique em <https://validar.it.gov.br>

Nome completo do professor-orientador: André Luiz Alves