



**UNIVERSIDADE FEDERAL DO CEARÁ**  
**FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE**  
**PROGRAMA DE ECONOMIA PROFISSIONAL**

**RICARDO DA SILVA REIS**

**PREDIÇÃO E IDENTIFICAÇÃO DE EMPRESAS NOTEIRAS UTILIZANDO  
MACHINE LEARNING NA SECRETARIA DE FAZENDA DO CEARÁ**

**FORTALEZA**

**2024**

RICARDO DA SILVA REIS

PREDIÇÃO E IDENTIFICAÇÃO DE EMPRESAS NOTEIRAS UTILIZANDO  
MACHINE LEARNING NA SECRETARIA DE FAZENDA DO CEARÁ

Dissertação apresentada ao Programa de Economia Profissional da Universidade Federal do Ceará, como requisito parcial para a obtenção do grau de Mestre em Economia do Setor Público. Área de concentração: Mestrado economia do setor público.

Orientador: Professor Dr. Sérgio Aquino de Souza

FORTALEZA

2024

Dados Internacionais de Catalogação na Publicação  
Universidade Federal do Ceará  
Sistema de Bibliotecas  
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

---

R312p Reis, Ricardo da Silva.  
Predição e identificação de empresas noteiras utilizando machine learning na secretaria de fazenda do Ceará / Ricardo da Silva Reis. – 2024.  
85 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Mestrado Profissional em Economia do Setor Público, Fortaleza, 2024.  
Orientação: Prof. Dr. Sérgio Aquino de Souza.

1. Noteiras. 2. Fraude. 3. Sonegação. 4. ICMS. 5. Machine Learning. I. Título.

CDD 330

---

RICARDO DA SILVA REIS

PREDIÇÃO E IDENTIFICAÇÃO DE EMPRESAS NOTEIRAS UTILIZANDO  
MACHINE LEARNING NA SECRETARIA DE FAZENDA DO CEARÁ

Dissertação apresentada ao Programa de Economia Profissional da Universidade Federal do Ceará, como requisito parcial para a obtenção do grau de Mestre em Economia do Setor Público. Área de concentração: Mestrado economia do setor público.

Aprovada em 27/ 06/ 2024

BANCA EXAMINADORA

---

Prof. Dr. Sérgio Aquino de Souza (Orientador)

Universidade Federal do Ceará (DEA/CAEN/UFC)

---

Prof. Dr. Ricardo Brito Soares

Universidade Federal do Ceará (DA/CAEN/UFC)

---

Prof. Dr. Rafael Barros Barbosa

Universidade Federal do Ceará (DEA/UFC)

## **AGRADECIMENTOS**

Agradeço a Deus pelas oportunidades que tem me oferecido.

Agradeço a minha esposa Samara pelo amor e carinho durante o processo.

Agradeço à Secretaria da Fazenda do Estado do Ceará, em especial a Coordenadoria de Pesquisa e Análise Fiscal, nas pessoas do Coordenador Raimundo Glison Pinheiro de Oliveira e da Orientadora Francisca Helena Paixão, pelo apoio institucional fundamental para a concretização desse trabalho.

Agradeço ao colega Auditor Fernando Castro de Mesquita por todo o suporte técnico e instrumental fornecidos durante essa pesquisa, além das conversas que se tornavam verdadeiras aulas de Machine Learning.

Agradeço ao meu orientador, Professor Doutor Sérgio Aquino de Souza, pelas sugestões, avaliações e orientações essenciais para o desenvolvimento desse projeto de pesquisa.

## RESUMO

As empresas noteiras são criadas com o objetivo de realizar a emissão de documentos fiscais fraudulentos, que não correspondem a uma operação de circulação efetiva de mercadoria ou prestação real de serviço, visando a geração de créditos de ICMS indevidos para serem utilizados pelos destinatários dos documentos fiscais, que poderão utilizar esses créditos “podres” para compensar o valor do ICMS devido ao fisco estadual. Nesse sentido, o presente trabalho consiste numa revisão das principais experiências de utilização de modelos de Machine Learning para a identificação e previsão de empresas noteiras nas Administrações Tributárias, bem como das principais características do ICMS e das sistemáticas de tributação adotadas. O trabalho realiza também um estudo teórico dos principais elementos presentes em uma empresa noteira, sócios “laranjas” e “testas de ferro”, e classifica essas empresas noteiras em três tipos de acordo com o grau de complexidade. Em seguida, é realizado um estudo com cinco modelos de Machine Learning para classificação (Regressão Logística, KNN, Rede Neural, Random Forest e XGBoost), visando a identificação e previsão de empresas noteiras na Secretaria de Fazenda do Estado do Ceará. Por fim, é feita uma comparação entre as métricas de avaliação dos modelos para definir quais modelos obtiveram melhor resultado

**Palavras – Chave:** Noteiras; Fraude; Sonegação; ICMS; Machine Learning.

## ABSTRACT

*Noteiras* companies are created with the aim of issuing fraudulent tax documents, which do not correspond to an effective circulation of goods or actual provision of services, with the aim of generating undue ICMS credits to be used by the recipients of the tax documents., who will be able to use these “bad” credits to offset the value of ICMS owed to the state tax authorities. In this sense, the present work consists of a review of the main experiences of using Machine Learning models for identifying and forecasting *noteiras* companies in Tax Administrations, as well as the main characteristics of ICMS and the taxation systems adopted. The work also carries out a theoretical study of the main elements present in a *noteiras* company, “orange” and “front” partners, and classifies these *noteiras* companies into three types according to the degree of complexity. Next, a study is carried out with five Machine Learning models for classification (Logistic Regression, KNN, Neural Network, Random Forest and XGBoost), aiming to identify and predict *noteiras* companies in the Ceará State Finance Department. Finally, a comparison is made between the model evaluation metrics to define which models obtained the best results.

**Keywords:** Noteiras; Fraud; Tax evasion; ICMS; Machine Learning.

## LISTA DE ILUSTRAÇÕES

|   |    |
|---|----|
| Figura 1 – Ilustração do Princípio da Não Cumulatividade.....                                 | 18 |
| Figura 2 – Apuração do ICMS da Cadeia Comercial.....  | 18 |
| Figura 3 – Ilustração do Processo de Substituição Tributária.....                             | 19 |
| Figura 4 – Apuração do ICMS próprio e do ICMS substituição tributária.....                    | 19 |
| Figura 5 – Cadeia de operação comercial interestadual com substituição por carga líquida..... | 21 |
| Figura 6 – Transferência de créditos inidôneos de ICMS.....                                   | 24 |
| Figura 7 – Visão Geral das Etapas do KDD.....   | 29 |
| Figura 8 – Visão Geral do CRISP-DM.....   | 29 |
| Figura 9 – Gráfico da Função Sigmoide $\sigma(t)$ .....                                       | 31 |
| Figura 10 – Modelo pra KNN para a classificação para K=3 ou K=6.....                          | 32 |
| Figura 11 – Arquitetura de uma Rede Neural Artificial.....                                    | 34 |
| Figura 12 – Rede Perceptron de uma camada.....  | 35 |
| Figura 13 – Floresta Aleatória com Bagging.....   | 37 |
| Figura 14 – Características do XGBoost.....   | 38 |
| Figura 15 – Visão geral do modelo XGBoost.....  | 39 |
| Figura 16 – Distribuição das Empresas Noteiras por Regime de Recolhimento.....                | 42 |
| Figura 17 – Distribuição das Empresas não Noteiras por Regime de<br>Recolhimento.....         | 42 |
| Figura 18 – Distribuição das Empresas Noteiras por segmento econômico.....                    | 44 |
| Figura 19 – Distribuição das Empresas Não Noteiras por segmento econômico.....                | 44 |
| Figura 20 – Visão Geral do Pré – Processamento dos Dados.....                                 | 46 |



|   |    |
|---|----|
| Figura 21 – Modelo de Classificação Random Forest para a seleção das Features relevantes..... | 47 |
| Figura 22 – Matriz de Confusão do modelo Random Forest para 1 amostra.....                    | 49 |
| Quadro 1 – Características de Sócios “Laranjas” e Testa de Ferro.....                         | 23 |
| Quadro 2 – Resumo dos tipos de Noteiras.....  | 27 |

## LISTA DE TABELAS

|  |    |
|--|----|
| Tabela 1 – Distribuição das Empresas Noteiras por Regime de Recolhimento.....        | 41 |
| Tabela 2 – Distribuição das Empresas não Noteiras por Regime de<br>Recolhimento..... | 42 |
| Tabela 3 – Distribuição das Empresas Noteiras por segmento econômico.....            | 43 |
| Tabela 4 – Distribuição das Empresas não Noteiras por segmento econômico.....        | 44 |
| Tabela 5 – Distribuição das <i>features</i> monetárias por categoria.....            | 45 |
| Tabela 6 – Distribuição das <i>features</i> não monetárias por categoria.....        | 46 |
| Tabela 7 – Métricas de Desempenho para 1 amostra.....                                | 49 |
| Tabela 8 – Métricas de desempenho para 10 amostras.....                              | 51 |

## SUMÁRIO

|            |  |           |
|------------|--|-----------|
| <b>1</b>   | <b>INTRODUÇÃO.....</b>   | <b>11</b> |
| <b>2</b>   | <b>REVISÃO DE LITERATURA.....</b>                                      | <b>14</b> |
| <b>3</b>   | <b>REVISÃO TEÓRICA.....</b>  | <b>16</b> |
| <b>3.1</b> | <b>O imposto sobre circulação de mercadorias.....</b>                  | <b>16</b> |
| 3.1.1      | Princípio da não cumulatividade.....                                   | 16        |
| 3.1.2      | Substituição tributária e carga líquida no estado do Ceará.....        | 18        |
| <b>3.2</b> | <b>Empresas Noteiras .....</b>   | <b>21</b> |
| 3.2.1      | Uso de interpostas pessoas.....  | 22        |
| 3.2.2      | Transferências de créditos inidôneos.....                              | 23        |
| 3.2.3      | Sonegação fiscal do ICMS substituição tributária ou carga líquida..... | 24        |
| 3.2.4      | Tipos de empresas Noteiras.....  | 25        |
| 3.2.4.1    | Noteiras Tipo I .....  | 25        |
| 3.2.4.2    | Noteiras Tipo II .....   | 25        |
| 3.2.4.3    | Noteiras Tipo III .....  | 26        |
| <b>3.3</b> | <b>Ciência de Dados.....</b>   | <b>27</b> |
| 3.3.1      | O processo KDD.....  | 28        |
| 3.3.2      | O processo CRISP-DM.....   | 29        |
| <b>3.4</b> | <b>Machine Learning.....</b>   | <b>30</b> |
| 3.4.1      | Regressão Logística.....   | 30        |
| 3.4.2      | Modelo KNN.....  | 32        |
| 3.4.3      | Rede Neural .....  | 34        |
| 3.4.4      | Random Forest .....  | 36        |

|            |   |           |
|------------|---|-----------|
| 3.4.5      | XGBoost.....  | 38        |
| <b>4</b>   | <b>METODOLOGIA.....</b>                                 | <b>40</b> |
| <b>4.1</b> | <b>Coleta dos Dados.....</b>                            | <b>40</b> |
| <b>4.2</b> | <b>Análise Exploratória dos Dados.....</b>              | <b>41</b> |
| 4.2.1      | Regime de Recolhimento.....                             | 41        |
| 4.2.2      | Segmento Econômico.....                                 | 43        |
| <b>4.3</b> | <b>Seleção das Features .....</b>                       | <b>45</b> |
| <b>4.4</b> | <b>Pré-Processamento.....</b>                           | <b>46</b> |
| <b>5</b>   | <b>DISCUSSÃO DOS RESULTADOS.....</b>                    | <b>48</b> |
| <b>6</b>   | <b>CONCLUSÃO.....</b>                                   | <b>52</b> |
|            | <b>REFERÊNCIAS.....</b>                                 | <b>55</b> |
|            | <b>APÊNDICE A - ANÁLISE EXPLORATÓRIA DOS DADOS.....</b> | <b>59</b> |
|            | <b>APÊNDICE B – LISTAS DE FEATURES.....</b>             | <b>69</b> |
|            | <b>APÊNDICE C – MÉTRICAS DOS MODELOS AVALIADOS.....</b> | <b>77</b> |

## 1 INTRODUÇÃO

As Administrações Tributárias têm como um grande desafio o combate às denominadas empresas noteiras (fantasmas ou de fachada), que apesar de não ser um termo conceituado na ordem legal, podem ser definidas como empresas fraudulentas criadas como o objetivo de emitir notas fiscais inidôneas que não correspondem a operações de circulação de mercadorias ou prestações de serviços reais, seja para a transmissão de créditos indevidos de ICMS, ou, para a regularização de estoque de mercadorias (SEFAZ RS, 2023).

As empresas noteiras têm causado considerável prejuízo à arrecadação tributária dos Estados e da União, além de possibilitarem a concorrência desleal por parte dos beneficiários da fraude. Tais empresas inidôneas ainda podem ser utilizadas para a emissão de notas fiscais para acobertarem a entrada fiscal e contábil de mercadorias provenientes de furto ou roubo, ou, a circulação ilícita de produtos provenientes de contrabando ou descaminho. (SEFAZ SC, 2013)

A Receita Federal do Brasil juntamente com a Secretaria de Fazenda do Estado de São Paulo e órgãos parceiros deflagraram, em maio de 2024, a Operação Metalmorfose visando o combate a um esquema que utilizava empresas noteiras para a emissão de notas fiscais fraudulentas com o fito de simular operações comerciais de produtos de cobre e sucata, com montante de notas fiscais frias de R\$ 17 bilhões, no período de 2018 a 2020 (RECEITA FEDERAL, 2024).

A Secretaria de Estado de Fazenda de Minas Gerais, em 4 de julho de 2023, deu andamento à 6ª fase da Operação Sinergia que teve como alvo organização criminosa especializada na criação de empresas noteiras para a perpetração da sonegação de impostos, em específico no setor de metais recicláveis. O modus operandi consistia na abertura de empresas noteiras para serem abastecidas com estoque fictício de crédito de ICMS por outras empresas fraudulentas, lesando inclusive outras unidades federativas, com estimativa de prejuízo aos cofres públicos de R\$ 96 milhões (SEFA MG, 2023).

O combate à sonegação fiscal é uma das principais atribuições das Administrações Tributárias, portanto identificar as empresas noteiras e realizar uma

contenção de danos, responsabilizando as beneficiárias do esquema fraudulento é tarefa relevante.

Os órgãos de fiscalização tributária têm atuado no sentido da identificação e penalização dessas empresas noteiras, bem como na responsabilização das empresas beneficiárias desse esquema de sonegação fiscal e das pessoas físicas envolvidas. Entretanto, dificuldades de ordem técnica, bem como a escassez de recursos humanos e materiais impedem que a atuação da fiscalização seja eficiente frente a velocidade e ao volume com que essas empresas fraudadoras são inseridas no sistema tributário brasileiro.

A desburocratização no processo de constituição e abertura de uma empresa no Brasil trouxe como subproduto a possibilidade da abertura de empresa fraudulentas em um tempo menor, sem maiores controles e garantias por parte do Estado. Tal situação fez com que aumentasse o número de empresas noteiras, permitindo que elas sejam criadas para existir durante um intervalo curto de tempo e causando prejuízos financeiros ao Estado de difícil recuperação. (Oliveira, 2023)

O problema relatado requer, por parte da Administração Tributária, o uso de ferramentas que possibilitem acompanhar a velocidade e o volume com que as empresas noteiras são criadas. Nesse sentido, o uso das ferramentas tradicionais de auditoria fiscal mostra-se ineficazes para tal tarefa, o que impõe a necessidade da busca por soluções alternativas.

O uso da Inteligência Artificial, de suas ferramentas e técnicas é uma alternativa de solução, uma vez que, com base em um banco de dados pré-existente de empresas fraudulentas já identificadas, é possível buscar características que sejam relevantes nesse grupo. Nesse sentido, é viável elaborar modelos que permitam a predição de empresas noteiras, intensificando a velocidade e a contundência da atuação da fiscalização tributária.

O contexto fático descrito tem levado os fiscos a buscarem formas de se anteverem ao problema, tais como elencar características que sejam comuns às notas fiscais fraudadas e ao próprio cadastro das empresas. Entretanto, esse processo de análise e identificação, apesar de já ser feito com o auxílio do cruzamento de banco de dados, depende bastante da capacidade técnica e experiência do Auditor Fiscal.

O desenvolvimento da Ciência de Dados e da Inteligência Artificial possibilitou o surgimento de iniciativas, em algumas Secretarias de Fazenda, da utilização de métodos e técnicas de *Machine Learning* para a predição de empresas noteiras.

A Secretaria de Fazenda do Estado de Goiás é um exemplo dessas iniciativas, conforme a matéria “Fisco goiano usa inteligência artificial para identificar empresas fantasmas” veiculada no site JORNAL OPÇÃO, em abril de 2023, foi executado um projeto pioneiro na SEFAZ GO do uso de redes neurais artificiais para a identificação e combate às empresas fantasmas, permitindo ao órgão atuação mais firme.

O presente projeto de pesquisa tem como objetivo principal desenvolver um modelo de identificação e predição de empresas noteiras, utilizando as técnicas e métodos de *Machine Learning*, no âmbito da Secretaria da Fazenda do Ceará, visando auxiliar o fisco cearense a combater com tempestividade e eficácia o mencionado mecanismo de sonegação fiscal.

A consecução do objetivo principal deverá ser precedida de outros objetivos (específicos), tais como construir um banco de dados de empresas noteiras já identificadas pela Secretaria de Fazenda, por conseguinte, obter um conjunto de empresas rotuladas como não noteiras. Desse conjunto total de empresas (noteiras e não noteiras) é necessário realizar uma análise estatística descritiva dos dados cadastrais e das notas fiscais emitidas e destinadas.

A etapa seguinte é identificar *features* (características) relevantes entre os dados das empresas (noteiras e não noteiras) e suas notas fiscais (emitidas e destinadas). De posse das *features*, é necessário definir quais modelos de aprendizado de máquina podem ser aplicados ao problema de classificação, bem como estabelecer e aplicar métricas e indicadores que possibilitem aferir os desempenhos dos modelos, e comparar os resultados entre os modelos.

## 2 REVISÃO DE LITERATURA

A literatura sobre uso de inteligência artificial para identificação de empresas noteiras é bem escassa. Entretanto, existem alguns trabalhos realizados sobre o tema ou complementares.

Gomes (2023) realizou um trabalho de investigação do uso de técnicas de análise de dados e Aprendizagem de Máquina para a identificação da fraude fiscal estruturada praticada por empresas constituídas exclusivamente para emitir créditos indevidos de ICMS, no âmbito da Secretaria de Fazenda do Distrito Federal. Dentre os modelos preditivos trabalhados, destacou-se como melhor modelo o Random Forest Classifier na base Máximo, seguido pelo mesmo modelo na base Soma, e, por último, o Gradient Boosting na base Soma.

Pinto e Fávero (2022) realizaram a construção de uma rede neural artificial para identificação de empresas fraudulentas, no âmbito da Secretaria da Fazenda do Estado de Goiás, utilizando-se de características definidas como relevantes dos contribuintes. Algumas variáveis utilizadas foram: localização, tipo de atividade econômica, porte, área do estabelecimento, notas fiscais de compra. Os resultados do trabalho demonstraram a viabilidade do uso da inteligência artificial para identificação das empresas fantasmas, pois as variáveis elencadas conseguiram explicar a variável dependente.

Arguiló e Quadrelli (2022) desenvolveram um modelo de Lógica Nebulosa com o objetivo de identificar eventuais desvios de conduta de empresas fornecedoras que caracterizem fraudes na emissão de notas fiscais falsas para transferir créditos de ICMS para as supostas empresas compradoras. O projeto foi realizado na Secretaria de Fazenda do Estado do Rio de Janeiro, resultando em um percentual de acerto elevado, tornando-se uma importante ferramenta no combate à sonegação fiscal.

Oliveira e Santos (2020) desenvolveram na Secretaria de Fazenda da Bahia um Sistema de Identificação de Risco de Contribuintes utilizando Redes Neurais Artificiais no modelo Perceptron de múltiplas camadas com retroalimentação, treinada com o algoritmo de retropropagação de erro. Os dados utilizados foram obtidos das



declarações feitas pelos contribuintes, o modelo obteve um índice de acerto de 71% na identificação de potenciais sonegadores do ICMS.

Os projetos relatados foram executados com o uso de dados não abertos, ou seja, não podem ser usados e compartilhados livremente, necessitando-se ter um perfil de acesso ou autorização legal para manipulação dos dados. Entretanto, há na literatura um trabalho de identificação de empresas fantasmas com aplicação de técnicas de Inteligência Artificial usando somente dados abertos.

Xavier et al (2021) utilizaram para a realização do projeto de identificação de contribuintes com perfis de inadimplentes dados abertos e públicos disponibilizados pela Receita Federal e pelo Conselho Administrativo Tributário do Estado de Goiás, além de outros cadastros públicos. No trabalho realizado foram implementados três modelos com o uso de recursos Random Forest, Redes Neurais e Grafos, a eficácia do modelo foi estimada em 98% de acerto do perfil inadimplente.

## 3 REVISÃO TEÓRICA

### 3.1 O imposto sobre circulação de mercadorias

O ICMS (Imposto sobre Circulação de Mercadorias e Prestação de Serviços) é um tributo estadual que tem sua incidência sobre operações de circulação de mercadorias e prestação de serviços de transporte interestadual e intermunicipal, e de comunicação. Em linhas gerais, é um imposto com incidência sobre o consumo, em que o ônus econômico do tributo recai sobre o consumidor final. (ALEXANDRE, 2023)

A base de cálculo do ICMS é o valor da operação ou da prestação, e a alíquota é definida, nas operações internas, por cada estado e, nas operações interestaduais, por resolução do Senado Federal. Ressalta-se que o cálculo do imposto é feito “por dentro”, ou seja, o valor do imposto compõe a sua própria base de cálculo. (Carraza, 2022)

O ICMS é a principal fonte de arrecadação dos estados brasileiros, utilizado, por esses, para custear os serviços públicos de educação, saúde, segurança, infraestrutura e as demais obrigações legais. Além disso, a Constituição Federal determina que 25% do total da arrecadação ICMS deve ser destinado aos municípios.

O Balanço Geral do Estado do Ceará do ano de 2023 indica que, no período de janeiro a dezembro do mencionado ano, o ICMS apresentou-se como a receita tributária mais expressiva, contribuindo com 76,58% da Receita Tributária, em termos nominais, R\$ 17,05 bilhões dos ingressos do montante arrecadado no período. (SEFAZ CE, 2024)

#### **3.1.1 Princípio da Não Cumulatividade**

A legislação tributária permite, ao contribuinte do ICMS, abater sobre o valor do ICMS devido pelas operações ou prestações tributadas o imposto pago nas operações ou prestações anteriores da cadeia econômica, de tal modo, que efetivamente a base de cálculo do imposto seja o valor adicionado em cada operação ou prestação, conforme determina o Princípio Constitucional da Não – Cumulatividade. Tal sistemática de apuração tem por objetivo evitar o efeito cascata

da tributação, em que a base de cálculo seria “inflada” pelo ICMS já pago nas operações anteriores.

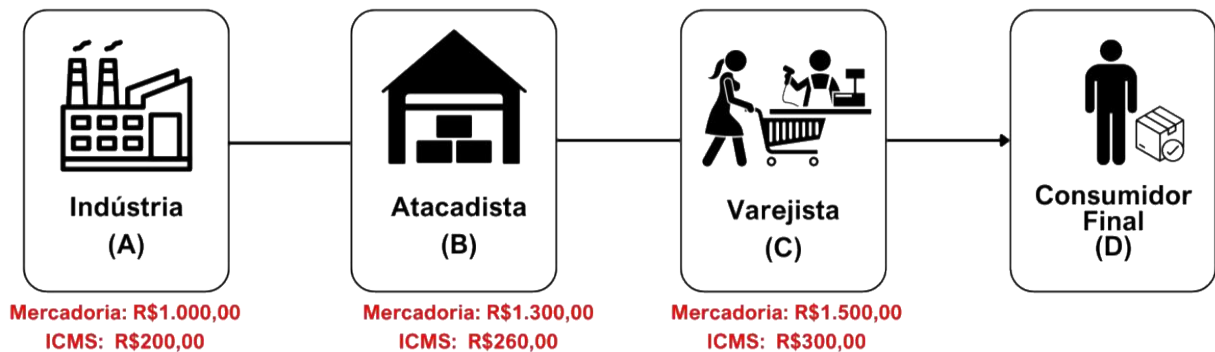
A sistemática básica de apuração do ICMS pelo método débito e crédito pode ser melhor elucidada com o seguinte exemplo. Suponha – se que um determinado estado da federação adote como alíquota interna modal de ICMS 20%, e que haja uma sucessão de operações comerciais com uma mercadoria desde o contribuinte industrial até o consumidor final.

Adota-se, para o exemplo, que uma indústria **A** realize uma operação interna de venda de uma mercadoria ao preço de R\$ 1.000,00 para uma distribuidora atacadista **B**. Nesse caso, a indústria **A** terá que pagar a título de ICMS para o estado em que ocorreu a operação, o valor de R\$ 200,00 ( $0,2 \times R\$ 1.000,00$ ), por outro lado, o atacadista **B** passa a ter um crédito de ICMS de R\$ 200,00.

A atacadista **B**, em um momento posterior, consiga revender a mercadoria para um comercial varejista **C**, pelo valor de R\$ 1.300,00. Nesse caso, a atacadista **B** teria que pagar a título de ICMS para o estado, o valor de R\$ 260,00 ( $0,2 \times R\$ 1.300,00$ ). Entretanto, como já dispunha de R\$ 200,00 de crédito da operação anterior, poderá utilizar esse valor para abater o valor do imposto devido e pagar somente R\$ 60,00 ( $R\$ 260,00 - R\$ 200,00$ ). Nessa situação concreta, surge o direito ao crédito de ICMS de R\$ 260,00 para comercial varejista **C** junto ao fisco estadual.

Por fim, o comercial varejista **C** realiza a revenda da mercadoria para um consumidor final **D**, ao preço de R\$ 1.500,00, portanto, surge o dever de pagar R\$ 300,00 correspondente ao ICMS da operação comercial. Entretanto, o varejista já dispõe de R\$ 260,00 de crédito de ICMS que pode ser utilizado para abater no valor do imposto devido, logo, deve recolher ao estado apenas R\$ 40,00 ( $R\$ 300 - R\$ 260,00$ ).

Figura 1- Ilustração do princípio da não cumulatividade.



Fonte: Elaborado pelo Autor

A análise geral da cadeia comercial demonstra que, após realizada as três operações comerciais, o valor a ser arrecadado pela Fazenda estadual é de R\$ 300,00 (R\$ 200,00 + R\$ 60,00 + R\$ 40,00), ou, em termos práticos 20% sobre o valor de venda final de R\$ 1.500,00. Tal sistemática, como já mencionado, evita o efeito cascata da tributação sobre o consumo, garantindo que o imposto incida apenas sobre o valor agregado em cada etapa da cadeia comercial, não inflando a base de cálculo do tributo.

Figura 2- Apuração do ICMS da Cadeia Comercial

| <br>IMPOSTO A RECOLHER - INDÚSTRIA (A) |  | <br>IMPOSTO A RECOLHER - ATACADISTA (B) |                | <br>IMPOSTO A RECOLHER - VAREJISTA (C) |                 |
|---|--|--|----------------|---|-----------------|
| R\$ 200,00 (I)  |  | R\$ 260,00 (II)  | R\$ 200,00 (I) | R\$ 300,00 (III)  | R\$ 260,00 (II) |
| R\$ 200,00  |  | R\$ 60,00  |                | R\$ 40,00   |                 |

Fonte: Elaborado pelo Autor

### 3.1.2 Substituição Tributária e Carga Líquida no Estado do Ceará

A legislação tributária do ICMS prevê a possibilidade de cobrança do tributo por meio da técnica da substituição tributária, que é a possibilidade de atribuir a responsabilidade legal pelo pagamento da carga tributária de toda a cadeia de operações a um único sujeito passivo da cadeia. Tal substituição pode ocorrer de três formas: regressiva, simultânea e progressiva.



A substituição tributária progressiva é a mais utilizada nas operações de circulação de mercadorias tributáveis pelos ICMS. Nesse tipo de substituição, em uma cadeia de operação comercial, o primeiro a realizar a circulação da mercadoria fica responsável por recolher o correspondente a carga tributária de toda a cadeia. No contexto do exemplo anterior, a indústria A seria responsável por recolher todo o ICMS da cadeia comercial, ou seja, em vez de recolher somente R\$ 200,00 referente a operação de venda para o atacadista B, terá que recolher R\$ 300,00 correspondente às operações subsequentes.

Figura 3 - Ilustração do processo de substituição tributária.



Fonte: Elaborado pelo Autor

Figura 4 - Apuração do ICMS próprio e do ICMS substituição tributária.

| <br><b>ICMS PRÓPRIO</b> | <br><b>ICMS SUBSTITUIÇÃO TRIBUTÁRIA</b> |
|--|--|
| R\$ 200,00 <sup>(I)</sup>  | R\$ 260,00 <sup>(II)</sup> R\$ 200,00 <sup>(I)</sup>   |
| R\$ 200,00   | R\$ 300,00      R\$ 260,00 <sup>(II)</sup>   |
| R\$ 200,00   | R\$ 100,00   |

Fonte: Elaborada pelo Autor

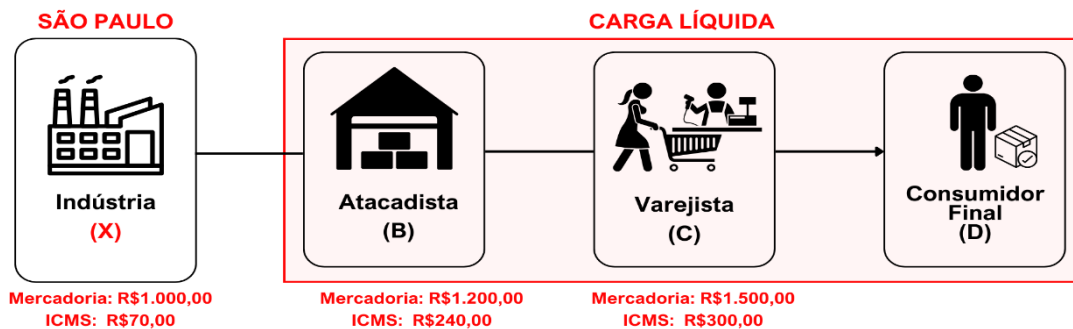
O estado do Ceará optou politicamente por adotar a substituição por carga líquida para contribuintes do ICMS que se enquadrem nas hipóteses elencadas na legislação tributária estadual e em suas regulamentações, de modo que os contribuintes cearenses ao adquirirem mercadorias de outros estados da federação passam a recolher o valor de ICMS correspondente às operações futuras da

mercadoria. Em síntese, a legislação estadual define uma alíquota que deve ser aplicada a base de cálculo da operação comercial, de tal forma, que já considere o crédito de ICMS correspondente, a alíquota de operação interestadual e interna, bem como a margem de valor agregado para as operações comerciais. Tal sistemática de tributação visa, conforme a própria norma, neutralizar a concorrência desleal entre contribuintes que exerçam a mesma atividade econômica e simplificar a tributação para os contribuintes

A sistemática de tributação por carga líquida pode ser exemplificada pela seguinte operação comercial tributável hipotética, uma indústria **X** do estado de São Paulo realiza a operação de venda de uma mercadoria pelo valor de R\$ 1.000 para um atacadista **B**, situado no estado do Ceará que posteriormente realiza a venda da mercadoria para uma empresa varejista **C** do mesmo estado pelo valor de R\$ 1.200,00. A empresa varejista **C** realiza, em seguida, a revenda da mercadoria para o consumidor final **D** por R\$ 1.500,00.

A realização da apuração do imposto a ser pago pela sistemática de débito e crédito do ICMS, implicaria que a indústria **X** teria que recolher 7% (alíquota interestadual) do valor da venda para o atacadista **B**, ou seja, R\$ 70,00 deveriam ser arrecadados para o estado de São Paulo, em contrapartida, o atacadista **B** passa a dispor de R\$ 70,00 em crédito de ICMS. O atacadista **B** deveria recolher de ICMS para o estado do Ceará, 20% do valor da operação para o varejista **C**, ou seja, R\$ 240,00 com o abatimento do crédito de R\$ 70,00 da operação anterior, totalizando R\$ 170,00. Já o comercial varejista **C** deve recolher ao estado do Ceará 20% do valor de revenda para o consumidor final, R\$ 300,00, descontando-se R\$ 240,00 que faz jus a título de crédito da operação de compra, totalizando R\$ 60,00 a ser arrecado. Em resumo, em termos tributários, o estado de São Paulo arrecadará R\$ 70,00 e o estado do Ceará, R\$ 230,00 (R\$ 190,00 + R\$ 40,00).

Figura 5 – Cadeia de operação comercial interestadual com substituição por carga líquida



Fonte: Elaborado pelo Autor

A sistemática de tributação por carga líquida visa arrecadar o valor de R\$ 230,00 já na entrada da mercadoria no estado do Ceará, de modo que o atacadista **B** seria responsável por recolher o valor do ICMS de toda a cadeia comercial de forma antecipada. Desse modo, em relação às operações internas subsequentes de venda da mercadoria não deve ser cobrado ICMS, pois já foi recolhido. Portanto, a adoção de tal sistemática possibilita ao fisco cearense centralizar o recolhimento do ICMS no momento da entrada da mercadoria no estado do Ceará, dispensando a necessidade da realização de múltiplos cálculos e pagamentos do tributo nas etapas subsequentes da cadeia comercial, permitindo maior eficiência na arrecadação e promovendo a neutralidade da concorrência desleal para contribuintes do mesmo segmento internamente.

### 3.2 Empresas Noteiras

As empresas noteiras, no contexto da tributação estadual, são empresas fraudulentas constituídas com o objetivo de realizar a emissão de notas fiscais para terceiros que não correspondem a uma efetiva operação de circulação de mercadoria mediante a perpetração da fraude fiscal estruturada, com vistas a realizar a transferência de créditos inidôneos de ICMS para beneficiários do esquema. Tais créditos de ICMS podem ser utilizados pelos beneficiários do esquema para abater no valor do ICMS que deve ser pago, possibilitando a sonegação fiscal do tributo, seja pela redução, ou mesmo, supressão do valor devido. (Alcantra, 2023)

A Doutrina de Inteligência Fiscal, conforme Protocolo ICMS 66 de 2009, entende a fraude fiscal estruturada como um ilícito penal de estruturas complexas,

perpetradas por organizações criminosas. No âmbito das empresas noteiras, a abertura legal na Junta Comercial e no Cadastro de Contribuintes é feita, em algumas vezes, com a utilização de documentos falsificados e que não são capazes de produzir efeitos jurídicos, tais como Carteiras de identidades, Carteiras de Habilitação, comprovantes de residência (conta de água, energia elétrica, internet) inidôneos ou produzidos mediante edição em aplicativos.

### **3.2.1 Uso de interpostas pessoas**

Há, por vezes, na abertura da noteira, a utilização, no quadro societário, de interpostas pessoas, popularmente denominados de “laranjas” ou “testas de ferro” com o objetivo de ocultar os verdadeiros sócios da empresa, visando evitar a responsabilização tributária e criminal e promover a blindagem patrimonial dos mesmos.

No caso dos sócios “laranjas” podem ser pessoas que aceitam fornecer seus dados em troca de uma remuneração financeira, ou mesmo, pessoas que tiveram seus dados pessoais obtidos de forma ilegal (vazamento de cadastros, golpes, furtos), tais pessoas geralmente tem uma participação passiva na prática delituosa, por vezes, até a desconhecendo. Já a figura do “testa de ferro” cede seus documentos pessoais mediante remuneração financeira e tem participação ativa na fraude, realizando a administração e representação da empresa noteira, mediante orientações e instruções dos verdadeiros donos do negócio. (Redação Conjur, 2017)



**Quadro 1** – Características de Sócios “Laranjas” e Testa de Ferro

| <b>Característica</b>         | <b>Sócios 'Laranjas'</b>  | <b>Testas de Ferro</b>   |
|-------------------------------|---|--|
| <i>Definição</i>              | Indivíduos cujos dados pessoais <u>são usados</u> para figurarem como sócios de empresas fraudulentas | Indivíduos que <u>cedem</u> seus dados pessoais para figurarem como sócios de empresas fraudulentas. |
| <i>Grau de Conhecimento</i>   | Inexistente/ inexpressivo   | Elevado  |
| <i>Grau de Atividade</i>      | Passiva   | Ativa  |
| <i>Compensação Financeira</i> | Sim /Não  | Sim  |
| <i>Tipos de Envolvimento</i>  | Inexistente/inexpressivo  | Elevado  |

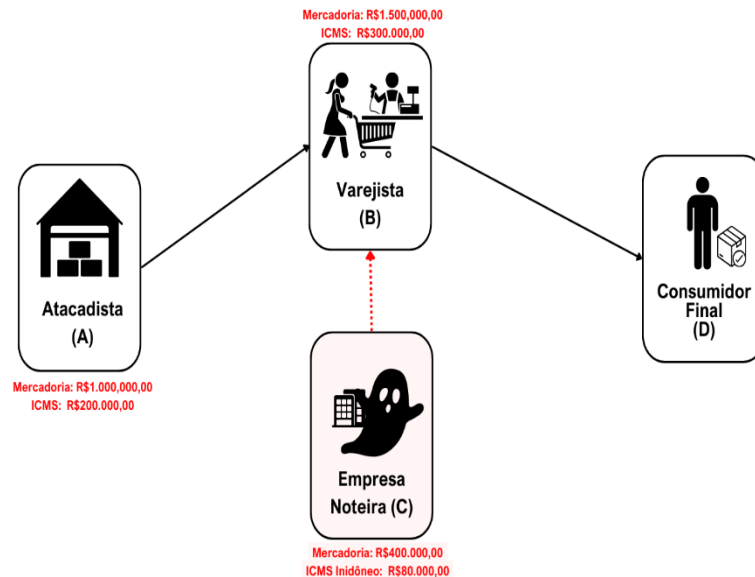
Fonte: Elaborado pelo Autor a partir das informações da Redação Conjur (2017)

### **3.2.2 Transferências de créditos inidôneos**

A compreensão da atuação das empresas noteiras para a transmissão de crédito inidôneos de ICMS pode ser feita a partir do seguinte exemplo, supondo-se que uma empresa atacadista **A** realize vendas internas para uma empresa varejista **B**, em um valor total de R\$ 1.000.000,00, com destaque de valor de ICMS de R\$ 200.000,00, que será o valor de débito tributário a ser pago pela atacadista, e, o valor do crédito a ser apropriado pelo varejista **B**. Desse modo, ao realizar as vendas para consumidor final, no total de R\$ 1.500.000,00, com valor de ICMS de R\$ 300.000,00, devendo recolher, ao fisco estadual, a título de ICMS, a diferença entre os débitos e os créditos, ou seja, R\$ 100.000,00 (R\$ 300.000,00 – R\$ 200.000,00). Entretanto, a empresa varejista **B**, visando reduzir o pagamento do tributo, faz uso do “serviço” de uma empresa noteira C que simula operações de venda, no valor de R\$ 400.000, para a empresa varejista **B**, que passa a dispor de um crédito (inidôneo) de ICMS de R\$ 80.000,00 (20% de R\$ 400.000,00). Assim, o valor de ICMS a ser recolhido pela empresa B passa a ser apenas R\$ 20.000 (R\$ 300.000,00 - R\$ 200.000,00 – R\$ 80.000,00), sonogando-se R\$ 80.000,00 em ICMS. Por fim, a empresa noteira **C**

passa a dever ao fisco estadual R\$ 80.000,00 em tributos, porém nem a noteira e nem os seus sócios possuem capacidade econômica para arcar com o pagamento, tornando-se irrecuperável o crédito tributário.

Figura 6 - Transferência de créditos inidôneos de ICMS



Fonte: Elaborado pelo Autor

### 3.2.3 Sonegação fiscal do ICMS substituição tributária ou carga líquida

As empresas noteiras também podem ser utilizadas para realizar a internalização no estado de mercadorias sem o recolhimento do ICMS substituição na entrada ou carga líquida. A título de exemplo, suponha uma empresa atacadista **A**, situada em outro estado do Nordeste, realiza operações de venda de mercadorias, no valor de R\$ 1.000.000,00 para a empresa varejista **B**, situada no estado do Ceará, e que tais mercadorias se enquadram nas hipóteses legais da tributação de carga líquida, com uma alíquota hipotética de 10%. Nesse tipo de operação, a empresa atacadista A deve recolher a título de ICMS para o estado em que está situado 12% do valor da operação, ou seja, R\$ 120.000,00 (12% de R\$ 1.000.000,00), e a empresa B deve recolher 10% do valor das mercadorias, R\$ 100.000,00 (10% de R\$ 1.000.000,00), para o estado do Ceará, correspondente a tributação de toda a cadeia subsequente, portanto, não realizando também destaque de ICMS para as vendas internas. Entretanto, buscando realizar a sonegação da tributação carga líquida, a varejista B passa a se valer do “serviço” fraudulento de uma empresa noteira **C**. Desse

modo, é realizado a simulação das aquisições interestaduais pela noteira **C**, situada no estado do Ceará, ou seja, para todos os feitos legais, a empresa atacadista **A** realiza a operação de venda para a empresa noteira **C** que passa a dever o valor do ICMS carga líquida, no nosso exemplo, R\$ 100.000,00, no entanto, nem a empresa noteira **C** e nem seus sócios possuem capacidade econômica para arcar com o pagamento do tributo, tornando o crédito tributário irrecuperável. Por fim, a empresa noteira **C** emite um documento fiscal simulando uma operação de venda interna para a empresa varejista **B**, que pode comercializar as mercadorias sem realizar o pagamento do tributo devido, uma vez, que “em tese” o tributo já foi retido pela noteira **C**.

### **3.2.4 Tipos de empresas noteiras**

As noteiras podem ser classificadas de acordo com o nível de complexidade de organização para a perpetração da fraude fiscal estruturada, tal classificação, apesar de não ser encontrada na literatura sobre a temática, é adotada empiricamente na realização nos trabalhos de identificação dessas empresas fraudulentas baseados nos padrões observados pelas autoridades fiscais.

#### **3.2.4.1 Noteiras Tipo I**

A classificação como Noteira do Tipo I é adotada para as empresas fraudulentas constituídas apenas formalmente, que não existem fisicamente, não existindo estabelecimento comercial no endereço informado no Cadastro Geral da Fazenda, de modo que por vezes o endereço corresponde a um imóvel residencial ou sequer existe fisicamente.

A realização de diligência *in loco*, na maioria das vezes, é suficiente para a autoridade fiscal constatar a inexistência de fato da empresa e dar provimento aos procedimentos administrativos cabíveis para a baixa da empresa e consequente interrupção das práticas fraudulentas.

#### **3.2.4.2 Noteiras Tipo II**

A classificação como Noteira Tipo II é adota para empresas fraudulentas que possuem existência física. Entretanto, o porte do estabelecimento comercial não condiz com o volume das operações comerciais supostamente realizadas, possuem estoque físico de mercadorias inexpressivo, ou, até mesmo, inexistente.

A análise das notas fiscais emitidas e destinadas demonstra divergência considerável entre os volumes de entradas e saídas, não correspondência entre a natureza dos produtos adquiridos e os produtos comercializados. Entretanto, com o objetivo de ludibriar a fiscalização tributária, realizam algumas operações comerciais reais, com ínfimo recolhimento do tributo.

A identificação desse tipo de empresa fraudulenta requer, da autoridade fiscal, a análise das informações fiscais e contábeis contidos nos sistemas corporativos da Administração Tributária, realização de diligência *in loco* e comprovação efetiva da não realização das operações comerciais junto aos fornecedores e aos clientes da empresa – circularização -, buscando verificar a inexistência de lastro financeiro, a não circulação física da mercadoria e demais elementos que atestem a simulação das operações comerciais.

#### **3.2.4.3 Noteiras Tipo III**

A empresa Noteira Tipo III possui uma estrutura de fraude fiscal estruturada mais complexa que a Noteira Tipo II, com estabelecimento comercial com porte físico compatível com o volume das operações comerciais. Há correspondência entre as entradas e as saídas de mercadorias, os produtos comercializados são da mesma natureza dos adquiridos, existindo, até mesmo lastro financeiro.

A identificação desse tipo de empresa fraudadora é muito complexa e demanda, por parte da fiscalização, além da análise das informações dos sistemas corporativos e da realização de visita *in loco*, o acesso, por meio de autorização judicial, aos dados das movimentações e transações bancárias dos envolvidos na fraude. Em casos mais graves, é necessário, em cooperação com a Polícia e o Ministério Público, a realização de operação de busca e apreensão, para a coleta do dado negado e de evidências que levem à responsabilização legal dos beneficiários.

**Quadro 2 – Resumo dos tipos de Noteiras.**

| <b>Tipo de Noteira</b> | <b>Descrição</b>   | <b>Características</b>  | <b>Métodos de Identificação</b>   |
|------------------------|--|---|---|
| Tipo I                 | Constituídas apenas formalmente.   | Não existem fisicamente;<br>Endereço inexistente ou corresponde a imóvel residencial.   | Diligência in loco.   |
| Tipo II                | Constituídas formalmente;<br>Estrutura física precária.  | Estabelecimento físico incompatível.<br>Estoque de mercadorias inexistente ou ínfimo.<br>Divergências entre o volume de entradas e saídas   | Análises fiscais e contábeis;<br>Diligência in loco;<br>Circularização junto aos fornecedores e clientes                                    |
| Tipo III               | Possuem estrutura física complexa;<br>Compatibilidade entre o volume de entradas e saídas;<br>Lastro financeiro. | Estabelecimento com porte para o volume das operações;<br>Simula o lastro financeiro das operações de venda e compra;<br>Realiza realmente parte operações comerciais declaradas. | Análise de informações fiscais e contábeis;<br>Diligência in loco;<br>Acesso judicial a dados bancários;<br>Operações de busca e apreensão. |

Fonte: Elaborado pelo Autor

### 3.3 Ciência de dados

A Ciência de Dados, segundo a International Business Machine Corporation – IBM (2024), é um campo de estudo multidisciplinar que congrega conhecimentos das áreas da Matemática, Estatística, Ciência da Computação, visando extrair conhecimento e insights dos dados coletados, mediante o emprego de técnicas estatísticas, mineração de dados, aprendizado de máquina e análise de big data, para o auxílio na tomada de decisões e na resolução de problemas complexos.

Segundo a Amazon Web Services – AWS (2024), a Ciência de Dados, pode ser utilizada para quatro tipos de abordagens:

- **Descritiva:** é o tipo mais básico de análise de dados, visa resumir e organizar os dados para descrever o comportamento passado dos dados e buscar padrões e

insights. É caracterizada por visualização gráfica como gráficos de pizza, de barras, de linhas;

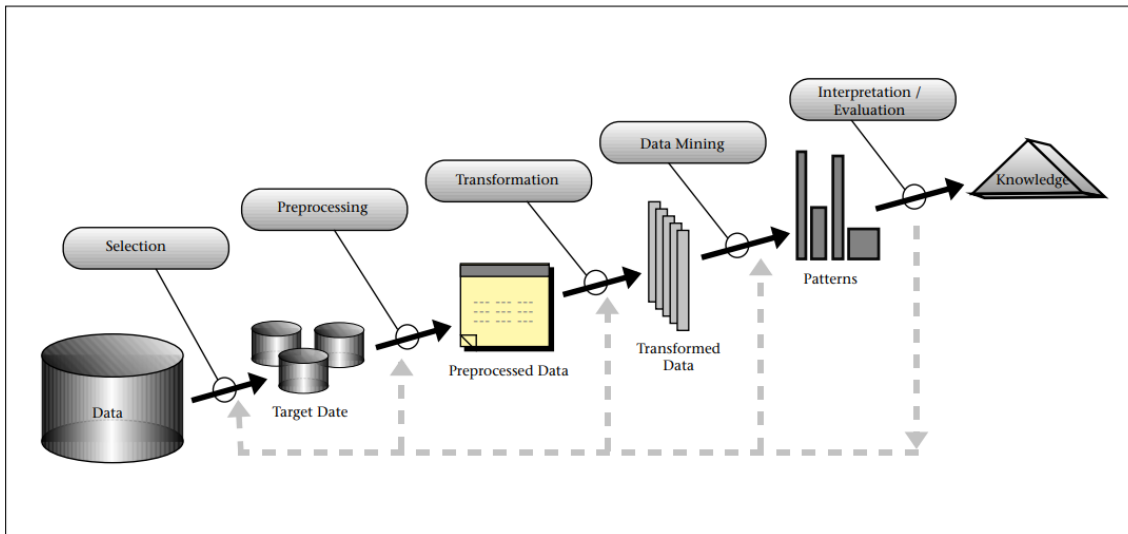
- Diagnóstica: é um tipo mais profundo e detalhado de análise de dados, busca compreender o porquê de os dados terem se comportado de determinado modo, faz uso de técnicas como drill-down, descoberta de dados, mineração de dados e correlações. Assim, os dados podem sofrer várias operações e transformações para a descoberta de novos padrões;
- Preditiva: os dados históricos são utilizados para realizar a previsão do comportamento futuro dos dados, faz o uso de técnicas como Machine Learning, previsão, correspondência de padrões e modelagem preditiva;
- Prescritiva: Esse tipo de análise não só realiza a previsão como também indica uma resposta ideal para o problema. Demonstra as implicações e potenciais de diferentes cenários e recomenda as melhores escolhas para a tomada de decisão. Faz uso de análise de gráficos, simulação, processamento de eventos complexos, redes neurais e mecanismos de recomendação de Machine Learning.

### **3.3.1 O processo KDD**

O processo KDD, sigla de Knowledge Discovery in Databases (Descoberta de conhecimento em banco de dados), é utilizado para extrair conhecimento dos dados, tal abordagem visa a descoberta de novas informações, possuindo várias fases, sendo as principais, segundo Fayyad et.al (1996):

- Seleção de Dados: Escolher os dados a serem trabalhados.
- Pré-processamento de Dados: Realizar a limpeza e tratamento dos dados ausentes, nulos ou outliers
- Transformação de Dados: Realizar processos como normalização, agregação e outras transformações nos dados, de modo a colocá-los sob um mesmo formato.
- Mineração de Dados: Utilizar técnicas, métodos ou modelos para a extração de informações e padrões úteis dos dados.
- Interpretação e Avaliação: Realizar o processo de validação dos resultados, verificando se os padrões identificados são suficientes ou podem ser melhorados.

Figura 7 – Visão Geral das Etapas do KDD

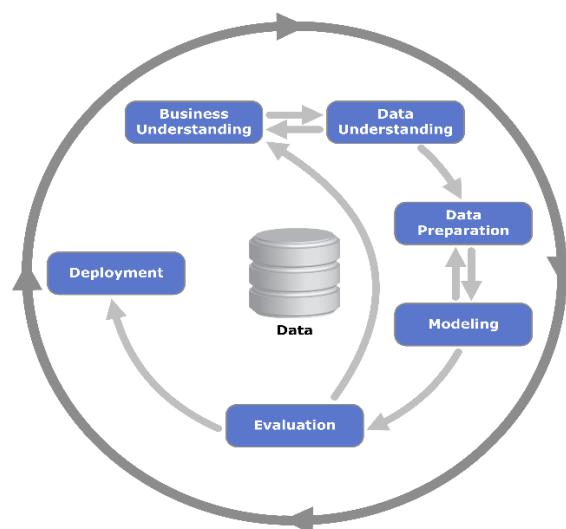


Fonte: (Fayyad, 1996)

### 3.3.2 O processo CRISP-DM

A sigla CRISP-DM, segundo Shearer (2000), abrevia a expressão Cross Industry Standard Process for Data Mining, que pode ser traduzida como Processo Padrão Inter-Indústrias para a Mineração de Dados, é uma metodologia que pode ser adotada em projetos de Ciência de Dados, considerada uma melhoria do KDD.

Figura 8 – Visão Geral do CRISP-DM



Fonte: Shearer (2000)

A CRISP-DM possui, segundo Shearer (2000), ao menos 6 fases básicas:

1. Entendimento do Negócio: É a fase em que se busca conhecer o negócio, as particularidades do projeto, as especificidades técnicas, os objetivos e os principais desafios envolvidos. Desse modo, é fundamental ter algum especialista do negócio envolvido para que possa auxiliar no processo;
2. Entendimento dos Dados: É a fase em que os dados são coletados, desenvolve-se familiaridade com os dados e realiza uma análise da qualidade destes, buscando alguns insights iniciais ou hipóteses não triviais;
3. Preparação dos Dados: É a fase em que, a partir dos dados brutos, se constrói o conjunto de dados final que alimentará o modelo. Envolve processos de seleção, limpeza, transformação, integração e formatação de dados;
4. Modelagem: É a fase que diversos modelos e técnicas são selecionados e aplicados ao conjunto de dados finais, ajustando-se os parâmetros iniciais. Pode ser dividida em seleção da técnica de modelagem, geração do projeto teste, criação de modelos e avaliação de modelos;
5. Avaliação: É a fase em que se avalia detalhadamente o modelo, revisando a construção do modelo para garantir que ele atinge aos propósitos estabelecidos;
6. Implantação: É a fase em que o modelo, após ser testado e validado, é incorporado ao processo de tomada de decisão.

### **3.4 Machine Learning**

Machine Learning ou Aprendizado de Máquina é uma subárea da Inteligência Artificial que visa a construção de algoritmos, modelos e técnicas que possam fazer com que os computadores aprendem com os dados para fazer estimativas ou tomar decisões ótimas. Desse modo, não é necessário definir previamente uma série de regras e instruções a serem seguidas pela máquina, pois o aprendizado será feito a partir de padrões identificados nos dados. (Géron, 2021)

#### **3.4.1 Regressão Logística**

O modelo de Regressão Logística é uma técnica de Machine Learning utilizada para a classificação de variáveis binárias ( $Y = 0,1$ ), baseada em aprendizado supervisionado. Este modelo permite que, a partir de um conjunto de amostras binárias associadas às suas *features* (características), realizar o treinamento de um modelo que permita separar as novas amostras.

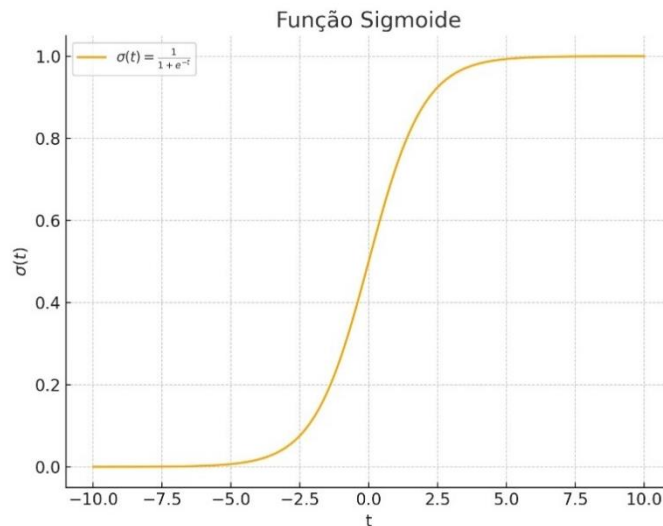
A função logística  $\sigma(t)$  ou função sigmoide é expressa por:



$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

onde  $t = \beta_0 + \beta_1 * X$ , sendo  $\beta_0$  o intercepto; e  $\beta_1$  coeficiente da variável dependente  $X$  e  $\sigma(t)$  determina a probabilidade de a amostra pertencer a classe positiva ( $Y=1$ ).

Figura 9 – Gráfico da Função Sigmoidal  $\sigma(t)$



Fonte: Elaborado pelo Autor

A aplicação de uma transformação logit sobre a relação entre as probabilidades de ocorrência dos eventos positivos ( $Y=1$ ) e negativo ( $Y=0$ ), é expressa por uma combinação linear do intercepto e dos coeficientes das variáveis dependentes:

$$\text{logit} \left( \frac{\sigma(t)}{1 - \sigma(t)} \right) = \beta_0 + \beta_1 * X$$

A estimativa dos coeficientes  $\beta_0$  e  $\beta_1$  é feita utilizando o método de máxima verossimilhança que busca definir os valores dos coeficientes que otimizam as probabilidades. O intercepto  $\beta_0$  indica a probabilidade do evento ( $Y=1$ ) ocorrer quando todas as variáveis preditoras forem igual a zero. Já o coeficiente  $\beta_1$  representa a variação do logit da probabilidade para variação de uma unidade na variável preditora.

Segundo Remigio (2020), o modelo de Regressão Logística possui as seguintes vantagens:

- Não está limitado a ser apenas um classificador, pois fornece também o valor da probabilidade para cada instância classificada;

- É um modelo de fácil implementação, rápido e de excelente desempenho, especialmente se os dados forem linearmente separáveis;
- Possui boa explicabilidade, uma vez que, a partir dos valores dos coeficientes, é possível identificar quais variáveis foram mais relevantes para o modelo e também a direção de associação com a variável predita ( $Y=0$  ou  $Y=1$ )

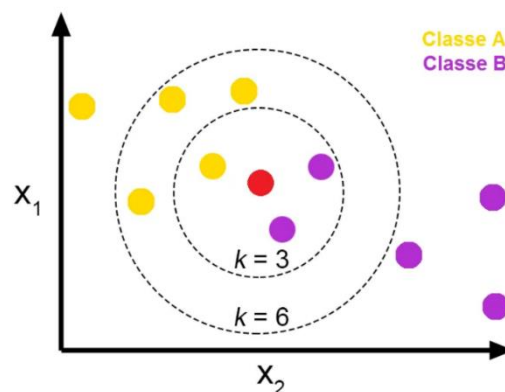
Remigio (2020) também aponta algumas desvantagens do modelo de Regressão Logística, são elas:

- O modelo pode ter problema com overfitting quando submetido a datasets de alta dimensionalidade, sendo recomendável a utilização de técnicas de regularização para atenuar o problema;
- Necessita de maior atenção na fase de pré-processamento dos dados comparado a outros modelos de classificação;
- Caso uma variável possua muito mais peso na decisão dos modelo que as demais, pode não haver convergência do coeficiente associado a essa variável.

### 3.4.2 Modelo KNN

O modelo K-Nearest Neighbors (KNN) é um modelo de Machine Learning baseado em aprendizado supervisionado que pode ser utilizado tanto para a classificação de dados com rótulos discretos quanto para a regressão de dados com rótulos contínuos. O fundamento do modelo é encontrar um número K de amostras rotuladas mais próximas da amostra a ser rotulada, realizado a previsão do rótulo a partir dessas. (BISHOP, 2006)

Figura 10 - Modelo pra KNN para a classificação para K=3 ou K=6



Fonte: Medium (2018)

Segundo a documentação do Scikit-learn (2024), o número de amostras  $K$  pode ser constante e definida pelo projetista ou variar de acordo com um raio de distância previamente definido. A distância adotada no modelo, via de regra, é a distância euclidiana  $d$ , mas é possível adotar outros padrões de distância:

$$d = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Algumas vantagens do modelo de aprendizado KNN (Bishop, 2006):

- É um modelo de fácil implementação e entendimento;
- Pode ser utilizado tanto para tarefas de classificação quanto para regressão;
- O modelo fornece resultado interpretativo, pois as previsões feitas são baseadas nos rótulos dos vizinhos mais próximos;
- O KNN é capaz de capturar as relações não lineares, pois não faz suposições sobre o limite de decisão;
- Pode ser utilizado para um número grande de problemas, uma vez que o modelo não faz suposições sobre a distribuição dos dados;
- O KNN não constrói modelo por problema, memoriza os dados de treinamento e os utiliza na previsão.

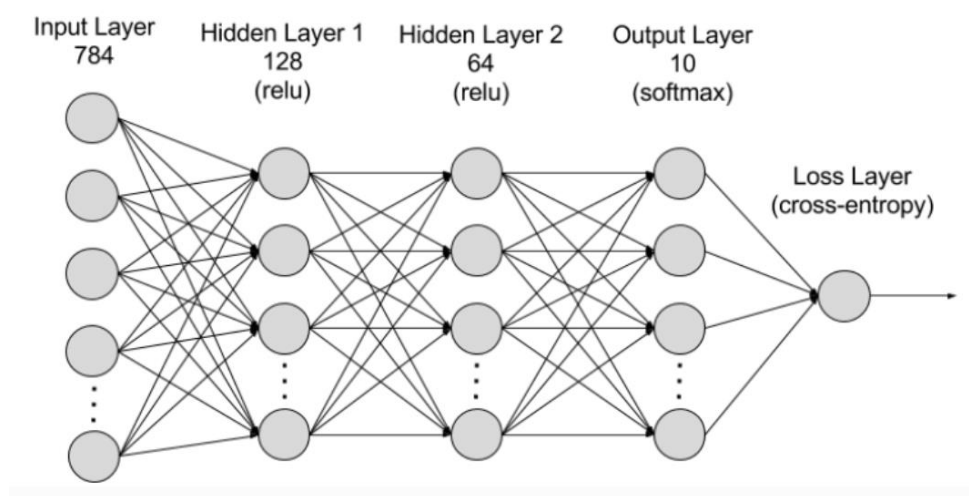
Algumas desvantagens do modelo de aprendizado KNN:

- Possui elevado custo computacional e de memória para o conjunto de grande dados e complexos;
- Tem desempenho reduzido para dados desequilibrados, possuindo maior tendência para os dados em maior quantidade;
- Alta sensibilidade a ruído para vizinhos mais próximos não representativos;
- Não é adequado para dados de alta dimensão, pois pode tornar a distância entre os pontos semelhantes;
- Definir o melhor valor para  $K$  pode ser um processo demorado;
- O método é sensível a outliers já que escolhe vizinhos com base na métrica de distância;
- Não se comporta bem para dados com valores ausentes.

### 3.4.3 Rede Neural

As Redes Neurais Artificiais são técnicas computacionais com um modelo matemático de inspiração nos neurônios dos seres vivos capazes de obterem conhecimento através das experiências. As redes neurais artificiais podem ser compostas por várias unidades de processamento, neurônios, conectadas por canais com pesos, sendo assim, cada unidade de processamento realiza operações matemáticas e os resultados são somados. (Haykin, 2001).

Figura 11 - Arquitetura de uma Rede Neural Artificial



Fonte: AWS (2024)

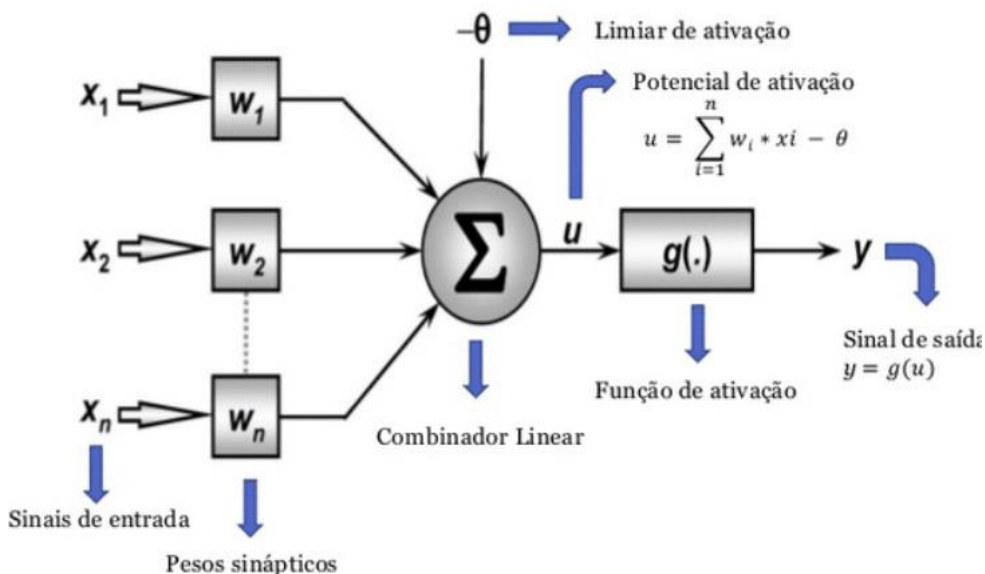
A arquitetura de uma Rede Neural simples pode ser dividida em três partes:

- A camada de entrada composto pelos neurônios que recebem os dados de entrada. Nessa camada, os nós representam as características dos dados;
- As camadas ocultas são as camadas que se encontram entre a camada de entrada e a camada de saída. Os nós, em uma camada oculta, são alimentadas por entradas ponderadas dos nós da camada imediatamente anterior, aplicando uma função de ativação e transmitindo o resultado para a camada imediatamente posterior. As funções de ativação normalmente utilizadas são a ReLU ( Redified Linear Unit), Sigmoid e Tanh;
- A camada de saída é formada pelos neurônios que produzem o resultado final do modelo. A quantidade de neurônios na camada de saída depende do problema

que se deseja resolver.

A perceptron é considerada a arquitetura mais básica das redes neurais modernas, tem como pai Frank Rosenblatt, em 1957. O seu funcionamento tem por base um neurônio artificial chamado de unidade lógica de limiar (TLU), ou, uma unidade de limiar linear (LTU). A TLU realiza a soma ponderada de suas entradas  $X_1, X_2, \dots, X_n$  pelos seus respectivos pesos,  $W_1, W_2, \dots, W_n$ , e se o seu resultado exceder o limiar de ativação, gera uma saída classificada como positiva. (GÉRON, 2021)

Figura 12 - Rede Perceptron de uma camada



Fonte: EMBARCADOS (2016)

O funcionamento de uma rede neural pode ser do tipo Forward Propagation quando os dados de entrada propagam de camada para camada até a camada de saída e, dessa forma, os neurônios vão calculando a soma ponderada das suas entradas e aplicando a função de ativação. Outro tipo de funcionamento é a Backward Propagation que consiste em ajustar os pesos das conexões entre os neurônios como o objetivo de otimizar o erro entre a saída prevista pela rede neural e a saída real. (GÉRON, 2021)

As Redes Neurais possuem as seguintes vantagens (Ribeiro, 2020):

- Capacidade de aprender padrões complexos e não lineares em dados;

- Possuem robustez a presença de ruídos ou dados incompletos;
- Capacidade de generalizar e fazer previsões para novos dados;
- Processamento dos dados em tempo real.

As Redes Neurais possuem as seguintes limitações (RIBEIRO, 2020):

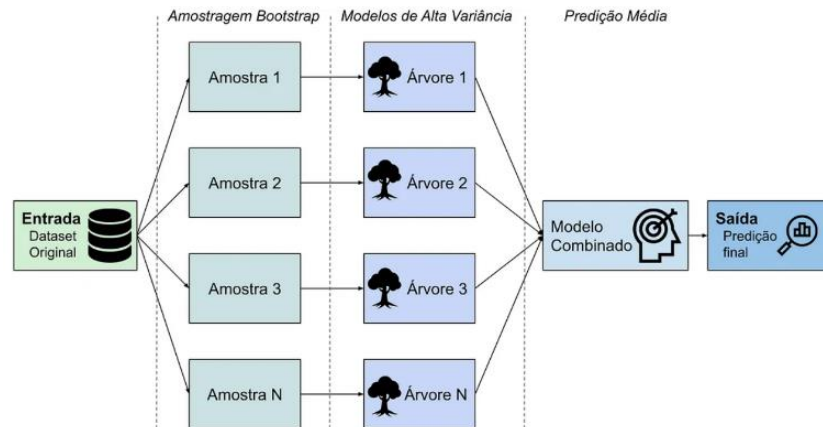
- Baixa interpretabilidade dos resultados;
- Necessidade de grandes quantidades de dados para realizar o treinamento;
- Tempo de treinamento relativamente alto do modelo;
- Dificuldade em escolher a arquitetura apropriada

#### **3.4.4 Random Forest**

Random Forest (Floresta Aleatória) é o modelo de Machine Learning que realiza a combinação das saídas de várias Árvores de Decisão para obter um único resultado, objetivando a melhoria da precisão e a redução do risco de overfitting, possuindo aplicabilidade tanto para problemas de classificação quanto para problemas de regressão. (EBAC, 2024)

O agrupamento de árvores de decisão geralmente é treinado pelo método Bagging (Bootstrap Aggregating) ou Pasting. No método Bagging, a amostragem é feita com reposição, com cada árvore sendo treinada por um subconjunto aleatório dos dados de treinos, podendo alguns dados serem repetidos ou não, tal abordagem diminui a variância e o overfitting do modelo. No método Pasting, a amostragem é feita sem reposição, cada árvore é treinada por um subconjunto aleatório dos dados de treinos, não havendo repetição dos dados, é viável para grande volume de dados. (Géron, 2020)

Figura 13 - Floresta Aleatória com Bagging



Fonte: Medium (2021)

Segundo a IBM (2024), o modelo de Florestas Aleatórias possui as seguintes vantagens:

- Reduzido risco de overfitting: árvores de decisão possuem elevado risco de overfitting, pois buscam ajustar todas as amostras dentro dos dados de treinamento. Entretanto, como há uma quantidade considerável de árvores de decisão em uma floresta aleatória, o modelo não irá sobre ajustar os dados, pois ao realizar a votação das classificações, reduz a variância e o erro de predição
- Flexibilidade: o modelo pode ser utilizado tanto para tarefas de classificação quanto de regressão, além do bagging otimizar os resultados do modelo;
- Importância das variáveis: o modelo possui alguns indicadores (Gini, por exemplo) que permitem estimar a importância de cada uma das variáveis no processo de predição.

As desvantagens do modelo Floresta Aleatória (IBM, 2024):

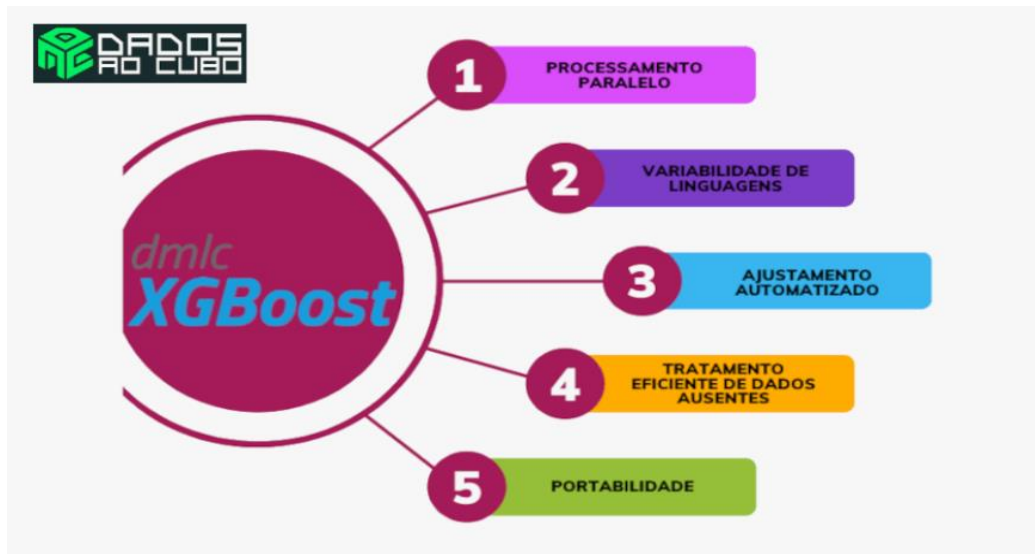
- Tempo elevado de processamento: o modelo irá calcular previsões para cada uma das árvores e realizar um processo de escolha da melhor classificação, isso demanda mais processamento computacional;
- Requer mais recursos: Necessitam de mais recursos para armazenar os resultados obtidos na predição;

- Baixa Interpretabilidade: possui uma maior complexidade do que as árvores de decisão para a compreensão da importância das variáveis no modelo.

### 3.4.5 XGBoost

O método XGBoost (Extreme Gradient Boosting) é um modelo complexo de Machine Learning baseado em Árvores de Decisão com gradient boosting conhecido pela sua eficiência e precisão. (Bruce & Bruce, 2019)

Figura 14 – Características do XGBoost

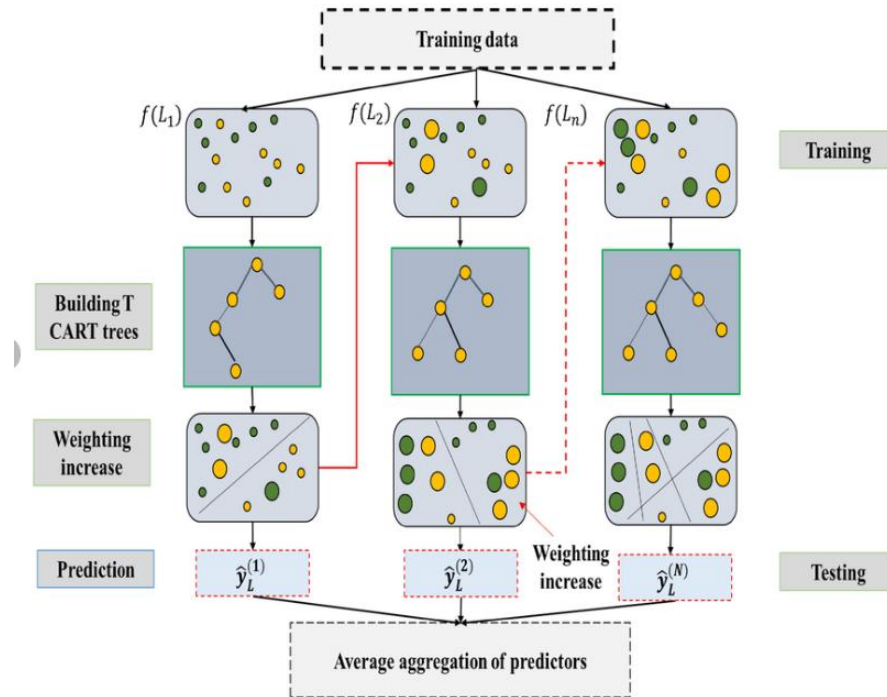


Fonte: Júnior (2022)

O XGBoost realiza a combinação de várias árvores fracas para criar uma árvore forte e fazer previsões precisas, sendo capaz de lidar com variáveis com tipos diversos de dados, variáveis categóricas ou numéricas. (Bruce & Bruce, 2019)



Figura 15 – Visão geral do modelo XGBoost



Fonte: Fonte: (Ali; Burhan, 2023)

## **4. METODOLOGIA**

A pesquisa a ser realizada quanto à abordagem será uma pesquisa quantitativa, haja vista que serão utilizados banco de dados, métricas e indicadores de desempenho e outros elementos quantificáveis do objeto a ser pesquisado (Lakatos, 1992).

O projeto visa aplicar métodos e técnicas já consolidados de aprendizado supervisionado de Machine Learning com vistas a identificar empresas noteiras e combater a sonegação fiscal do ICMS, é, portanto, uma pesquisa, quanto à natureza, aplicada.

O presente trabalho visa realizar um estudo focado nos contribuintes do ICMS do Estado do Ceará, considerando as particularidades da legislação estadual, bem como os registros anteriores de empresas consideradas noteiras pela Administração Pública. Nesse sentido, é possível dizer que, quanto aos objetivos definidos, é uma pesquisa exploratória de um estudo de caso. (Gil, 1991)

Os procedimentos que serão adotados para a execução do projeto, tais como construção de um banco de dados, análise estatística dos dados, realização de modelagem e testagem do modelo, são típicas de uma pesquisa experimental.

### **4.1 Coleta dos Dados**

A coleta dos dados inicialmente consistiu no levantamento, nos diversos bancos de dados e sistemas corporativos da Secretaria de Fazenda do Estado do Ceará, de empresas contribuintes do ICMS que tenham sido identificadas e comprovadas como empresas noteiras. Desse modo, foi construída uma lista com um total de 101 empresas noteiras que seriam utilizadas como amostras positivas para treino e teste dos modelos de Machine Learning.

A coleta posterior consistiu na consulta, nos bancos de dados e aos sistemas corporativos da SEFAZ CE, de empresas contribuintes do ICMS que tenham sido fiscalizadas ou monitoradas com indícios de serem empresas noteiras e tenham sido identificadas e constatadas como empresas não noteiras. Entretanto, não foi possível obter nenhuma amostra de empresa não noteira, dada a inexistência de registros históricos com essas condicionantes estabelecidas.

A construção dos modelos de classificação por meio de aprendizado supervisionado requer a existência de amostras prévias das classes a serem separadas, no caso concreto, das classes binárias - noteiras ou não noteiras. Nesse sentido, como forma de superar essa limitação, foi adotada a hipótese de que dentre o universo de empresas contribuintes de ICMS do estado do Ceará, o percentual de empresas noteiras representa algo inferior a 3%, de tal forma, que a seleção aleatória de um subconjunto das empresas existentes, retornaria amostras satisfatórias de empresa não noteiras. Portanto, foram coletadas de forma aleatória 600 empresas para a compor a lista das empresas não noteiras.

## 4.2 Análise Exploratória dos Dados

O processo de Análise Exploratória dos Dados (EDA) tem por objetivo realizar o conhecimento dos dados em busca de padrões, detecção de anomalias, outliers ou insights que possam ser úteis no processo de *Feature Engineering* (Construção das características), valendo-se de representações gráficas.

### 4.2.1 Regime de Recolhimento

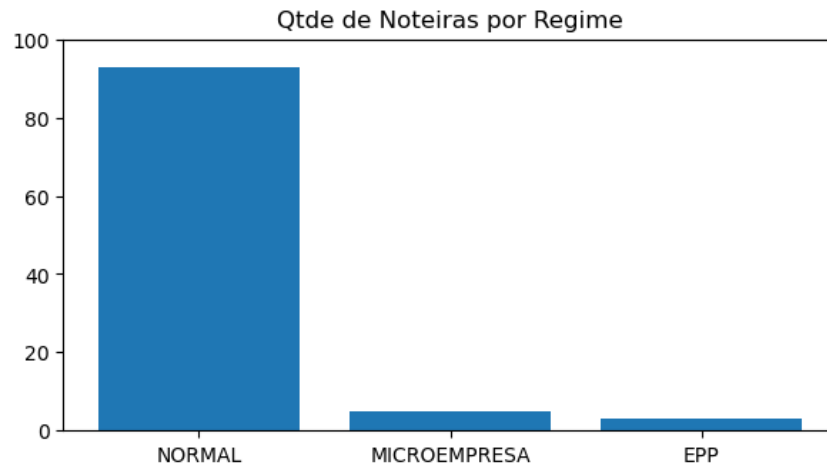
A verificação da distribuição das amostras das empresas noteiras por Regime de Recolhimento demonstrou que 93 empresas (92,08%) pertencem ao regime Normal de recolhimento, 5 noteiras (4,95%) são Microempresas, e as outras 3 (2,97%) são Empresas de Pequeno Porte.

Tabela 1 – Distribuição das Empresas Noteiras por Regime de Recolhimento

| Regime de Recolhimento | Qtde | Percentual (%) |
|------------------------|------|----------------|
| NORMAL                 | 93   | 92,08          |
| MICROEMPRESA           | 5    | 4,95           |
| EPP                    | 3    | 2,97           |

Fonte - Elaborado pelo Autor

Figura 16 – Distribuição das Empresas Noteiras por Regime de Recolhimento



Fonte - Elaborado pelo Autor

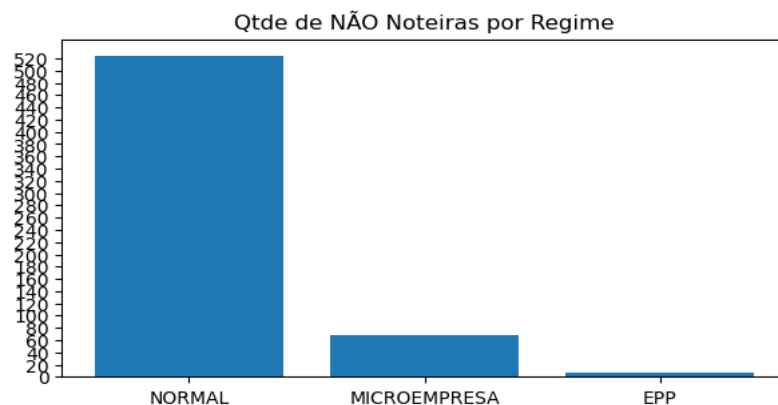
A análise da distribuição das amostras das empresas não noteiras por Regime de Recolhimento demonstrou que 525 empresas (87,5%) são do regime Normal, 69 (11,5%) são Microempresas e 6 (1%) são Empresas de Pequeno Porte.

Tabela 2 – Distribuição das Empresas não Noteiras por Regime de Recolhimento

| Regime de Recolhimento | Qtde | Percentual (%) |
|------------------------|------|----------------|
| NORMAL                 | 525  | 87,5           |
| MICROEMPRESA           | 69   | 11,5           |
| EPP                    | 6    | 1              |

Fonte – Elaborado pelo Autor

Figura 17– Distribuição das Empresas não Noteiras por Regime de Recolhimento



Fonte – Elaborado pelo Autor

A concentração de empresas noteiras nos regimes de recolhimento para Microempresas e Empresas de Pequeno Porte é inferior a 8% do total, já no caso das empresas não noteiras apenas 12,5% se concentram nesses regimes. Os referidos regimes estão contidos dentro do SIMPLES NACIONAL que possui um regramento jurídico diferenciado, permitindo a transferência de crédito de ICMS em valores bem menores que o regime Normal. Desse modo, considerando que o foco do trabalho é a construção de modelos de Machine Learning para a identificação e predição das empresas noteiras, torna-se uma opção técnica de projeto considerar apenas as 93 empresas noteiras pertencentes ao regime de recolhimento Normal para o treinamento e teste dos modelos.

#### **4.2.2 Segmento Econômico**

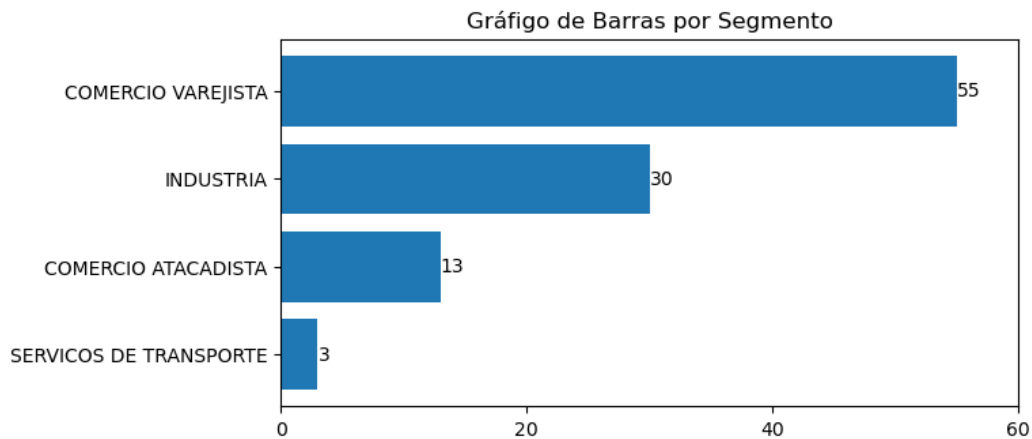
A análise da distribuição das empresas noteiras por Segmento Econômico demonstra que 55 empresas (54,5%) são Comércio Varejista, 30 noteiras (29,7%) são Indústrias, 13 (12,9%) são Comércio Atacadista e 3 (2,9%) são empresas do segmento de transporte.

Tabela 3 – Distribuição das Empresas Noteiras por segmento econômico

| <b>Segmento Econômico</b> | <b>Qtde</b> |
|---------------------------|-------------|
| COMERCIO VAREJISTA        | 55          |
| INDUSTRIA                 | 30          |
| COMERCIO ATACADISTA       | 13          |
| SERVICOS DE TRANSPORTE    | 3           |

Fonte – Elaborado pelo Autor

Figura 18 – Distribuição das Empresas Noteiras por segmento econômico



Fonte – Elaborado pelo Autor

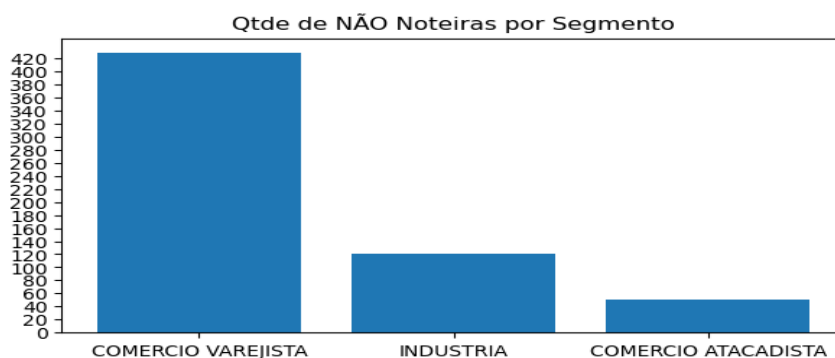
A distribuição das empresas não noteiras por Segmento Econômico demonstra que 429 empresas (71,5%) são comércios Varejistas, 120 empresas (20%) são Indústrias e 51 (8,5%) empresas são comércio Atacadista.

Tabela 4 – Distribuição das Empresas Não Noteiras por segmento econômico

| Segmento Econômico  | Qtde |
|---------------------|------|
| COMERCIO VAREJISTA  | 429  |
| INDUSTRIA           | 120  |
| COMERCIO ATACADISTA | 51   |

Fonte – Elaborado pelo Autor

Figura 19 – Distribuição das Empresas Não Noteiras por segmento econômico



Fonte – Elaborado pelo Autor

As empresas noteiras do segmento de transporte correspondem a 2,9% do conjunto total e as não noteiras não possuem empresas do segmento de transporte. Portanto, optou-se por não considerar, para o treinamento e teste do modelo, as amostras das empresas noteiras do segmento de transporte, restando o total de 91 amostras de empresas não noteiras, pois 2 das empresas noteiras do segmento de transportes também são do regime de recolhimento normal.

Um outro argumento para a exclusão das empresas do segmento de transporte é o fato de o documento fiscal que acoberta a prestação de serviços de transportes, Conhecimento de Transporte Eletrônico (CT-e), possuir características diferentes da Nota Fiscal Eletrônica (NF-e) que acoberta as operações de circulação de mercadorias dos outros segmentos econômicos (Varejo, Atacado e Industrial).

### 4.3 Seleção das Features

A seleção das *features* utilizadas na construção dos modelos de Machine Learning teve como parâmetro os resultados obtidos no processo de Análise Exploratória dos Dados e na etapa de Entendimento do Negócio. Assim foram construídas 107 *features*, que foram divididas em dois tipos: 55 *features* quantitativas monetárias e 52 *features* quantitativas não monetárias.

As *features* do tipo quantitativas monetárias são todas aquelas que podem ser expressas em termos monetários, e foram subdivididas em 5 categorias, conforme o atributo ou fonte dos dados utilizados: Capital Social, Débitos, Arrecadação, Notas Fiscais Emitidas, Nota Fiscais Recebidas.

Tabela 5 – Distribuição das *features* monetárias por categoria

| CATEGORIA               | QUANTIDADE |
|-------------------------|------------|
| CAPITAL SOCIAL          | 2          |
| ARRECADAÇÃO             | 12         |
| DÉBITOS                 | 13         |
| NOTAS FISCAIS RECEBIDAS | 14         |
| NOTAS FISCAS EMITIDAS   | 14         |
| <b>TOTAL</b>            | <b>55</b>  |

Fonte – Elaborado pelo Autor

As *features* do tipo quantitativas não monetárias são todas aquelas não podem ser expressas em termos monetários, e foram subdivididas em 7 categorias: Quantidades, Notas Fiscais, Destinatários/Fornecedores, Percentuais de Notas

Fiscais, Atividades e Mudanças, Percentuais de Atividades do Contador e Valores Acumulados.

Tabela 6 – Distribuição das *features* não monetárias por categoria

| CATEGORIA                             | QUANTIDADE |
|---------------------------------------|------------|
| QUANTIDADES (CGF E CNPJ)              | 12         |
| NOTAS FISCAIS                         | 10         |
| DESTINATÁRIOS/FORNECEDORES            | 10         |
| PERCENTUAIS DE NOTAS FISCAIS          | 8          |
| ATIVIDADES E MUDANÇAS                 | 7          |
| PERCENTUAIS DE ATIVIDADES DO CONTADOR | 3          |
| VALORES ACUMULADOS                    | 2          |
| <b>TOTAL</b>                          | <b>52</b>  |

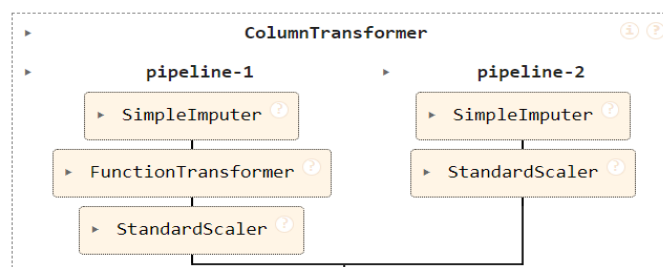
Fonte – Elaborado pelo Autor

#### 4.4 Pré-Processamento

O pré-processamento dos dados é a etapa em que é realizada a formatação dos dados, tratamento dos dados ausentes, nulos e outliers, bem como a normalização e padronização das variáveis.

O ColumnTransformer é uma ferramenta que aplica transformações diferentes para tipos diferentes de colunas nos dados. No pipeline -1 foram tratados os dados das features quantitativas monetárias e no pipeline – 2, as features quantitativas não monetárias.

Figura 20 – Visão Geral do Pré – Processamento dos Dados

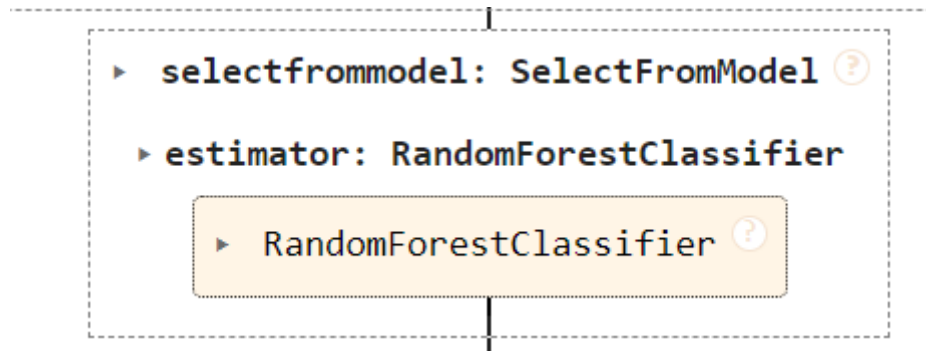


Fonte – Elaborado pelo Autor

A etapa seguinte foi utilizar o modelo de classificação Random Forest para selecionar, dentre as 107 *features* iniciais, aquelas que são mais relevantes para o treinamento do modelo. Desse modo, foram selecionadas, pelo modelo de classificação Random Forest, 30 *features*, das quais 17 são *features* quantitativas monetárias e 13 são *features* quantitativas não monetárias.



Figura 21 – Modelo de Classificação Random Forest para a seleção das Features relevantes



Fonte – Elaborado pelo Autor

## 5. DISCUSSÃO DOS RESULTADOS

A análise das métricas de avaliação para todos os modelos de classificação aplicados a 91 empresas noteiras e a uma amostra aleatória de 600 empresas não noteiras demonstra que todos os modelos avaliados obtiveram métricas de desempenho superiores a 0,93, quando se considera como referência a classe das não noteiras (Classe = 0), esse resultado é explicado pelo desbalanceamento entre a quantidade de exemplos de não noteiras (600) e quantidade de empresas noteiras (91). Entretanto, tais métricas de desempenho para a classe 0 não são relevantes para a avaliação dos modelos, haja vista que o foco do trabalho é a identificação e predição das empresas noteiras, representadas pela Classe 1, ou classe Positiva.

A análise e comparação das métricas de desempenho em relação a Classe 1, classe das empresas noteiras, demonstra que o modelo de classificação Random Forest é o que apresenta o melhor resultado para a previsão de empresas noteiras.

A Precisão do modelo Random Forest para a classe 1 é de 0,8953, o Recall ou Sensibilidade é de 0,8461, a média F1-score é 0,87, e, a Acurácia é de 0,9667, em todas essas métricas o modelo obteve resultado melhor que dos demais modelos.

O modelo de Regressão Logística teve o segundo melhor desempenho dentre os modelos com as seguintes métricas para a classe 1 (noteiras), Recall de 0,7582, f1-score de 0,7667, e Acurácia de 0,9392, já a Precisão foi de 0,7752 ligeiramente inferior a Precisão de 0,7831 do modelo XGBoost. Entretanto, é interessante pontuar que a Regressão Logística é um modelo bem mais simples que o XGBoost e possui uma alta explicabilidade do modelo através de seu intercepto e de seus coeficientes por *features* utilizadas.

O modelo que teve o desempenho pior foi o KNN com Precisão de 0,5913, Recall de 0,6044, f1-score de 0,5978 e Acurácia de 0,8929, tal fato pode ser explicado pela razão de o modelo KNN realizar a classificação dos dados não rotulados pelo critério de distância euclidiana dos K vizinhos mais próximos. Sendo assim, possui alto grau de viés quando as amostras de treinamento e teste se encontram desbalanceadas, que é o que ocorre no caso concreto (91/600), tendendo a classificar as amostras não rotuladas com a classe de maior quantidade. Além disso, esse modelo é considerado não generalizante uma vez que apenas “decora” os dados rotulados e estima a distância para os não rotulados.

Tabela 7 – Métricas de Desempenho para 1 amostra

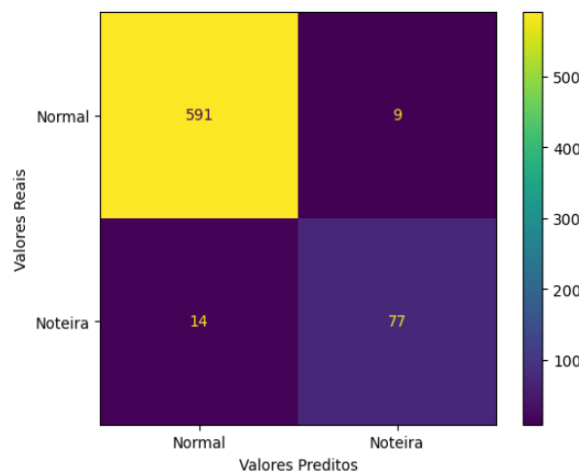
| Modelo              | Classe | precision | recall   | f1-score | accuracy | support |
|---------------------|--------|-----------|----------|----------|----------|---------|
| Regressao_Logistica | 0      | 0,963455  | 0,966667 | 0,965058 | 0,939219 | 600     |
| Regressao_Logistica | 1      | 0,775281  | 0,758242 | 0,766667 | 0,939219 | 91      |
| KNN                 | 0      | 0,939799  | 0,936667 | 0,93823  | 0,939219 | 600     |
| KNN                 | 1      | 0,591398  | 0,604396 | 0,597826 | 0,939219 | 91      |
| Rede_Neural         | 0      | 0,958541  | 0,963333 | 0,960931 | 0,939219 | 600     |
| Rede_Neural         | 1      | 0,75      | 0,725275 | 0,73743  | 0,939219 | 91      |
| Random_Forest       | 0      | 0,97686   | 0,985    | 0,980913 | 0,939219 | 600     |
| Random_Forest       | 1      | 0,895349  | 0,846154 | 0,870056 | 0,939219 | 91      |
| XGboost             | 0      | 0,957237  | 0,97     | 0,963576 | 0,939219 | 600     |
| XGboost             | 1      | 0,783133  | 0,714286 | 0,747126 | 0,939219 | 91      |

Fonte: Elaborado pelo Autor

A Matriz de Confusão do modelo de classificação Random Forest para 1 amostra demonstra que das 600 empresas não noteiras, o modelo conseguiu identificar 591 como não noteiras (Verdadeiros Negativos) e classificou erroneamente 9 empresas como noteiras (Falsos Positivos). Por outro lado, das 91 empresas noteiras, o modelo classificou corretamente 77 empresas como noteiras (Verdadeiros Positivos), e 14 como não noteiras (Falsos Negativos).

Figura 22 – Matriz de Confusão do modelo Random Forest para 1 amostra

Matriz de Confusão por CGF no Dataset de Teste



Fonte: Elaborado pelo Autor

As métricas de desempenho podem ser obtidas através dos valores contidos na Matriz de Confusão, para tal, tem -se  $VN = 591$ ,  $VP = 77$ ,  $FN = 14$ ,  $FP = 9$ :

- A Acurácia representa a proporção de previsões feitas corretamente pelo modelo em relação ao número total de previsões  $\frac{VN+VP}{VN+VP+FN+FP} = \frac{591+77}{591+77+14+9} = \frac{668}{691} = 0,9667$ . O valor de acurácia elevado indica que o modelo acerta muito, porém é necessário utilizar outras métricas para avaliar o modelo, pois em situações, como o do presente projeto, em que há amostras desbalanceadas o modelo pode estar prevendo apenas a classe que possui maior quantidade, não generalizando a classificação;
- A precisão representa a proporção das previsões feitas corretamente como positivas em relação ao valor total previsões como positivas  $\frac{VP}{VP+FP} = \frac{77}{77+9} = \frac{77}{86} = 0,8953$ . Essa métrica estima a qualidade das previsões positivas, porém não deve ser avaliada individualmente para a determinação do desempenho do modelo;
- O Recall ou Sensibilidade representa uma proporção das previsões positivas feitas corretamente em relação ao total de amostras positivas  $\frac{VP}{VP+FN} = \frac{77}{77+14} = \frac{77}{91} = 0,8461$ . É uma métrica que avalia a capacidade do modelo prever todas as amostras positivas existentes, não deve ser considerado individualmente para a determinação do desempenho do modelo;
- O f1-score é uma métrica que combina e equilibra os resultados do Recall e da Precisão  $2 \times \frac{PRECISÃO \cdot RECALL}{PRECISÃO + RECALL} = 2 \times \frac{0,8953 \cdot 0,8461}{0,8953 + 0,8461} = 0,87$ . É a melhor métrica para avaliar modelos com amostras desbalanceadas, que é o caso do presente projeto.

A análise e comparação das métricas dos modelos de classificação foi realizada para apenas 1 amostra de 91 noteiras e 600 empresas não noteiras selecionadas aleatoriamente, porém é possível que a amostra selecionada seja a pior amostra, ou mesmo, a melhor amostra dentre todas as existentes. Portanto, para reduzir o grau de viés na escolha da amostra, os modelos foram treinados para 10 amostras diferentes e calculada a média das métricas de desempenho.

O modelo Random Forest foi o modelo que obteve o melhor desempenho considerando a média das métricas para 10 amostras aleatórias, com f1-score de 0,8086, Precisão de 0,8310, Recall de 0,7879 e Acurácia de 0,9508. Esse resultado ratifica o resultado inicial para 1 amostra em que o Random Forest foi o melhor modelo.

O modelo de Regressão Logística, assim como no caso com 1 amostra, foi o segundo melhor modelo, com f1-score de 0,7465, Precisão de 0,7799, Recall de 0,7168 e Acurácia de 0,9360, mesmo sendo um método mais simples comparado ao XGBoost.

O modelo KNN, para 10 amostras, continuou sendo o modelo com o pior desempenho dentre os cinco modelos avaliados, com f1-score de 0,6353, Precisão de 0,6391, Recall de 0,6330 e Acurácia de 0,9041.

Tabela 8 – Métricas de desempenho para 10 amostras

| <b>Modelo</b>       | <b>F1-Score<br/>(Classe 0)</b> | <b>Precisão<br/>(Classe 0)</b> | <b>Recall<br/>(Classe 0)</b> | <b>F1-Score<br/>(Classe 1)</b> | <b>Precisão<br/>(Classe 1)</b> | <b>Recall<br/>(Classe 1)</b> | <b>Acurácia</b> |
|---------------------|--------------------------------|--------------------------------|------------------------------|--------------------------------|--------------------------------|------------------------------|-----------------|
| Regressão Logística | 0,963396                       | 0,957548                       | 0,969333                     | 0,746542                       | 0,779946                       | 0,716844                     | 0,936035        |
| KNN                 | 0,944844                       | 0,944405                       | 0,945333                     | 0,635299                       | 0,639093                       | 0,632967                     | 0,904197        |
| Rede Neural         | 0,957454                       | 0,952998                       | 0,962                        | 0,708539                       | 0,733677                       | 0,686813                     | 0,92576         |
| Random Forest       | 0,971768                       | 0,968081                       | 0,9755                       | 0,808572                       | 0,831035                       | 0,787912                     | 0,950796        |
| XGBoost             | 0,959506                       | 0,955421                       | 0,963667                     | 0,723797                       | 0,747131                       | 0,703297                     | 0,929378        |

Fonte: Elaborado pelo Autor

## 6. CONCLUSÃO

As empresas noteiras são empresas fraudulentas criadas como objetivo de realizar a emissão de documentos fiscais que não correspondam a uma efetiva operação de circulação de mercadorias ou prestação de serviços, visando gerar créditos inidôneos de ICMS para serem aproveitados pelos destinatários dos documentos fiscais na compensação do valor de imposto a ser pago ao fisco estadual. Além disso, as empresas noteiras podem ser utilizadas como destinatárias de documentos fiscais, em operações comerciais interestaduais, de modo a suportar a carga tributária da Substituição Tributária, Carga Líquida ou ICMS Antecipado, permitindo que as mercadorias circulem pelo Estado, sem o pagamento do tributo devido. Por fim, as empresas noteiras podem ainda emitir documentos fiscais para a regularização de estoque, “esquentar” mercadorias provenientes de roubo e contrabando, dos destinatários desses documentos fiscais.

As Administrações Tributárias têm como um dos seus principais objetivos o combate a esses tipos de fraudes, porém, diante da flexibilização para abertura de empresas, as empresas noteiras têm sido criadas em volumes cada vez maiores. Diante disso, o presente trabalho teve como foco a construção de modelos de Machine Learning para realizar a identificação e predição dessas empresas noteiras no contexto da Secretaria de Fazenda do Estado do Ceará, visando combater essas fraudes em uma velocidade maior do que as ferramentas tradicionais, mitigando a sonegação fiscal e a fraude fiscal estruturada.

A pesquisa inicialmente tratou de buscar experiências similares em outras Secretarias de Fazenda, bem como realizar o levantamento do Referencial Teórico existente sobre Ciência de Dados, Machine Learning, modelos de classificação, métricas de desempenho, bem como outras ferramentas teóricas necessárias. Assim, buscou-se delimitar o escopo do trabalho, através do Entendimento do Negócio, partindo logo em seguida para a Análise Exploratória dos Dados, com vistas a seleção e construção de features para serem utilizadas nos treinamentos e testes dos modelos.

A construção do banco de amostras de empresas noteiras baseou-se na lista de 101 empresas que já haviam sido identificadas pela Secretaria de Fazenda como empresas noteiras. Entretanto, por questões de projeto, foram consideradas

apenas empresas pertencentes ao Regime de Recolhimento Normal e Segmento Econômico Atacado, Varejo e Indústria, reduzindo para 91 amostras de empresas noteiras.

O banco das amostras das empresas não noteiras, em função da inexistência de registros anteriores, foi gerado de forma aleatória com 600 empresas do universo de contribuintes, pertencentes ao Regime de Recolhimento Normal e Segmento Econômico Atacado, Varejo e Indústria considerando a hipótese de que o percentual de empresas noteiras existentes é inferior a 3% do total.

O processo de seleção e construção de features produziu um total de 107 features, das quais 55 eram do tipo quantitativas monetárias e as outras 52, quantitativas não monetárias. Entretanto, durante o processo de pré-processamento dos dados, utilizando um modelo de classificação Random Forest, foram selecionadas desse total apenas 30 features, 17 quantitativas monetárias e 13 quantitativas não monetárias.

Os dados foram treinados e testados, em 5 modelos de classificação de Machine Learning, Regressão Logística, KNN, Random Forest, Rede Neural e XGBoost, com as 30 features, 91 amostras de empresas noteiras e 600 exemplos aleatórios de empresas não noteiras. Desse modo, considerando para apenas 1 amostra aleatória, o modelo de classificação que obteve o melhor desempenho foi o Random Forest, em todas as métricas de desempenho para a classe das empresas noteiras, f1-score, Precisão, Recall e Acurácia. Já o segundo melhor modelo foi a Regressão Logística, apesar de ser um modelo mais simples de classificação. Por fim, o modelo que obteve o pior desempenho foi o modelo KNN, em todas as métricas de desempenho.

Os modelos passaram por uma segunda avaliação, considerando dessa vez 10 amostras aleatórias em vez de apenas uma, e realizando a média das métricas de desempenho, o modelo de classificação Random Forest obteve novamente o melhor desempenho em todas as métricas de desempenho. O modelo de Regressão Logística obteve novamente o segundo melhor desempenho, e o modelo KNN obteve o pior desempenho dentre os 5 modelos avaliados.

Os resultados obtidos com o presente trabalho foram avaliados como satisfatórios, atendendo, portanto, aos objetivos inicialmente definidos.

Sugere-se para trabalhos futuros a construção de uma base de empresas não noteiras provenientes de avaliações e monitoramentos de Auditoria, bem como a consideração dos índices oficiais de inflação para atualização monetária das features quantitativas monetárias, otimização dos hiperparâmetros dos modelos de modo a obter ganhos nas métricas de desempenho.

A elaboração de um método, ou mesmo, um modelo que permita a explicabilidade das identificações e previsões realizadas pelos modelos de classificação é uma melhoria que poderia ser adotada em novos trabalhos.



## REFERÊNCIAS

AGUILÓ, Rafael R. QUADRELLI, Giovane. **APLICAÇÃO DA LÓGICA NEBULOSA NA DETECÇÃO DE FRAUDE DE ICMS**. REUCP, Petrópolis, RJ. Volume 16 n° 2(2022). P 53-63.

ALCANTARA, Alexandre. **Empresas noteiras: realidade também no Canadá**. Editoria: Prof. Alexandre Alcantara. 2 nov. 2023. Auditoria Fiscal, Crimes Tributários, Fraude Contábil, IVA (IBS & CBS). Disponível em: <https://alcantara.pro.br/portal/2023/11/02/empresas-noteiras-realidade-tambem-no-canada/>. Acesso em: 22 jun. 2024.

ALEXANDRE, Ricardo. **Direito Tributário**. 17. ed. São Paulo: Juspodivm, 2023. ISBN 9788544242510.

ALI, Zainab Hasan; BURHAN, Abbas M. **Hybrid machine learning approach for construction cost estimation: An evaluation of extreme gradient boosting model**. Asian Journal of Civil Engineering, v. 24, n. 7, p. 2427-2442, 2023.

AWS. **O que é uma rede neural?** Disponível em: <https://aws.amazon.com/pt/what-is/neural-network/>. Acesso em: 14 jun. 2024.

BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. New York: Springer Science+Business Media, LLC, 2006.

BRASIL. **Lei Complementar n. 87, de 13 de setembro de 1996**. Dispõe sobre o imposto dos Estados e do Distrito Federal sobre operações relativas à circulação de mercadorias e sobre prestações de serviços de transporte interestadual e intermunicipal e de comunicação, e dá outras providências. Diário Oficial da União: Brasília, DF, 16 set. 1996. Disponível em: [http://www.planalto.gov.br/ccivil\\_03/leis/lcp/lcp87.htm](http://www.planalto.gov.br/ccivil_03/leis/lcp/lcp87.htm). Acesso em: 11 jun. 2024.

BRASIL. Conselho Nacional de Política Fazendária. **Protocolo ICMS 66, de 3 de julho de 2009**. Dispõe sobre a instituição do Sistema de Inteligência Fiscal (SIF) e intercâmbio de informações entre as unidades da Federação. Disponível em: <[https://www.confaz.fazenda.gov.br/legislacao/protocolos/2009/pt066\\_09](https://www.confaz.fazenda.gov.br/legislacao/protocolos/2009/pt066_09)>. Acesso em: 22 jun. 2024.

BRUCE, Andrew; BRUCE, Peter. **Estatística Prática Para Cientistas de Dados: 50 Conceitos Essenciais**. 1ª ed. Alta Books, 2019.

CARRAZZA, Roque Antonio. **ICMS**. 19. ed. Salvador: Juspodivm, 2022. 800 p. ISBN 9786558600312.

DataCamp. **K-Nearest Neighbors (KNN) Classification with R Tutorial**. Disponível em: <https://www.datacamp.com/pt/tutorial/k-nearest-neighbors-knn-classification-with-r-tutorial>. Acesso em: 14 jun. 2024.

EMBARCADOS. **Rede Perceptron de uma Única Camada**. Disponível em:

<https://embarcados.com.br/rede-perceptron-de-uma-unica-camada/>. Acesso em: 14 jun. 2024.

Fisco goiano usa inteligência artificial para identificar empresas fantasmas. **Jornal Opção**, Goiania, ano 47, 25 abr. 2023. Tecnologia. Disponível em: < <https://www.jornalopcao.com.br/tecnologia/fisco-goiano-usa-inteligencia-artificial-para-identificar-empresas-fantasmas-485334/> >. Acesso em: 01 out. 2023.

FAYYAD, Usama; PIATETSKY-SHAPIO, Gregory; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery in Databases**. AI Magazine, v. 17, n. 3, p. 37-54, 1996.

Graphical scheme of XGBoost model. Disponível em: [https://www.researchgate.net/figure/Graphical-scheme-of-XGBoost-model\\_fig1\\_370000558](https://www.researchgate.net/figure/Graphical-scheme-of-XGBoost-model_fig1_370000558). Acesso em: 16 jun. 2024.

GÉRON, Aurélien. **Mãos à obra: aprendizado de máquina com Scikit-Learn, Keras & TensorFlow: conceitos, ferramentas e técnicas para a construção de sistemas inteligentes**. 1. ed. Rio de Janeiro: Alta Books, 2021. 640 p. ISBN 978-8550815480.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 2. ed. SP: Atlas, 1991.

GOMES, Gunther Siqueira Lemos. **Identificação de práticas de evasão fiscal utilizando aprendizagem de máquina: o caso das empresas de fachada e os créditos ilegais de ICMS**. 2023. 122 f. Dissertação (Mestrado em Governança, Tecnologia e Inovação) - Universidade Católica de Brasília, Brasília, 2023. Orientador: Prof. Dr. Remis Balaniuk.

Haykin, S. S. **Redes neurais: princípios e prática**. 2ª ed., Bookman, 2001.

JUNIOR, Onédio. **O Guia do XGBoost com Python**. Dados ao Cubo, 28 abr. 2022. Disponível em: <https://dadosaocubo.com/o-guia-do-xgboost-com-python/>. Acesso em: 16 jun. 2024.

KNN: K-Nearest Neighbors. Towards Data Science, Medium, 27 nov. 2018. Disponível em: <https://medium.com/towards-data-science/knn-k-nearest-neighbors-1-a4707b24bd1d>. Acesso em: 14 jun. 2024.

LAKATOS, Eva e Marconi, Marina. **Metodologia do Trabalho Científico**. SP : Atlas, 1992.

Mato Grosso integra operação nacional de combate às empresas noteiras. Secretaria de Fazenda do Mato Grosso, 2024. Disponível em: <https://www5.sefaz.mt.gov.br/-/10942780-mato-grosso-integra-operacao-nacional-de-combate-as-empresas-noteiras>. Acesso em: 11 jun. 2024.

MINAS GERAIS. Secretaria de Estado de Fazenda. 6ª fase da Operação Sinergia mira organização criminoso envolvida na criação de empresas fantasmas. 07 jul. 2023. Disponível em:

[https://www.fazenda.mg.gov.br/noticias/2023/2023.07.04\\_fase6sinergia.html](https://www.fazenda.mg.gov.br/noticias/2023/2023.07.04_fase6sinergia.html). Acesso em: 22 jun. 2024.

OLIVEIRA, Francisco N. SANTOS, Luis P.G. **Estratégias para Combater a Sonegação Fiscal**. GESTÃO, FINANÇAS E CONTABILIDADE, Salvador, BA. v.10. n.1. p.42-64,

OLIVEIRA, Marcelo Fernandes de. **Proposta de um modelo de combate a empresas noteiras**. 2023.

Operação Cuprum é deflagrada no setor metalmeccânico para combater a sonegação decorrente de empresas noteiras. **Secretaria de Fazenda do Rio Grande do Sul**, Porto Alegre, 06 abr. 2023. Disponível em:

<https://www.fazenda.rs.gov.br/conteudo/18517/operacao-cuprum-e-deflagrada-no-setor-metalmeccanico-para-combater-sonegacao-decorrente-de-empresas-noteiras>.

Acesso em: 01 out. 2023.

Operação Metalmorfose: Receita Federal e órgãos parceiros combatem esquema que emitiu R\$ 17 bilhões em notas fiscais frias. **Receita Federal**, 2024. Disponível em:

<https://www.gov.br/receitafederal/pt-br/assuntos/noticias/2024/maio/operacao-metalmorfose-receita-federal-e-orgaos-parceiros-combatem-esquema-que-emitiu-r-17-bilhoes-em-notas-fiscais-frias>. Acesso em: 11 jun. 2024.

Operação Nota Branca desarticula esquema de sonegação fiscal. Secretaria de Estado da Fazenda de Santa Catarina, 2013. Disponível em:

<https://www.sef.sc.gov.br/noticias/operacao-nota-branca-desarticula-esquema-de-sonegacao-fiscal>. Acesso em: 11 jun. 2024.

Operação “Expresso” desmantela esquema bilionário de sonegação; prejuízo no PR chega a R\$ 100 milhões. Secretaria de Fazenda do Paraná, Curitiba, 16 mar. 2023. Disponível em: <https://www.fazenda.pr.gov.br/Noticia/Operacao-Expresso-desmantela-esquema-bilionario-de-sonegacao-prejuizo-no-PR-chega-R-100>. Acesso em: 11 maio 2024.

O que é ciência de dados? IBM, 2024. Disponível em: <https://www.ibm.com/br-pt/topics/data-science>. Acesso em: 11 jun. 2024.

PEREIRA, Andreza Priscila. **Profissionalização da figura do laranja nas sociedades empresariais**. Conjur, 22 abr. 2024. Disponível em:

<https://www.conjur.com.br/2024-abr-22/profissionalizacao-da-figura-do-laranja-nas-sociedades-empresariais/>. Acesso em: 11 jun. 2024.

Pinto, Ricardo Costa. Fávero, Patrícia Belfiora. **Aplicação de rede neural artificial para auxiliar a fiscalização tributária na identificação de empresas noteiras**. 2022. Monografia(MBA em Data Science e Analytics). Universidade de São Paulo

REDAÇÃO CONJUR. **Juiz diferencia "testa de ferro" de "laranja" ao condenar acusado**. Conjur, 03 jun. 2017. Disponível em: <https://www.conjur.com.br/2017-jun-03/juiz-diferencia-testa-ferro-laranja-condenar-acusado/>. Acesso em: 11 jun. 2024.

REMIGIO, M. S. **Regressão Logística (Logistic Regression)**. Medium, 17 ago. 2020. Disponível em: <https://medium.com/@msremigio/regress%C3%A3o-log%C3%ADstica-logistic-regression-997c6259ff9a>. Acesso em: 14 jun. 2024.

RIBEIRO, Thiago. **Fundamentos das Redes Neurais: Teoria e Prática**. Medium, 26 abr. 2020. Disponível em: <https://medium.com/@thiago2002sr/fundamentos-das-redes-neurais-teoria-e-pr%C3%A1tica-056afdee06dd>. Acesso em: 14 jun. 2024.

SECRETARIA DA FAZENDA DO CEARÁ. **Balanco Geral do Estado do Ceará - 2023**. Disponível em: <https://www.sefaz.ce.gov.br/2024/04/08/sefaz-ceara-publica-balanco-geral-do-estado-de-2023/>. Acesso em: 11 jun. 2024.

SHEARER, Colin. **The CRISP-DM Model: The New Blueprint for Data Mining**. *Journal of Data Warehousing*, v. 5, n. 4, p. 13-22, 2000. Scikit-learn: Machine Learning in Python. **Nearest Neighbors**. Disponível em: <https://scikit-learn.org/stable/modules/neighbors.html#>. Acesso em: 14 jun. 2024.

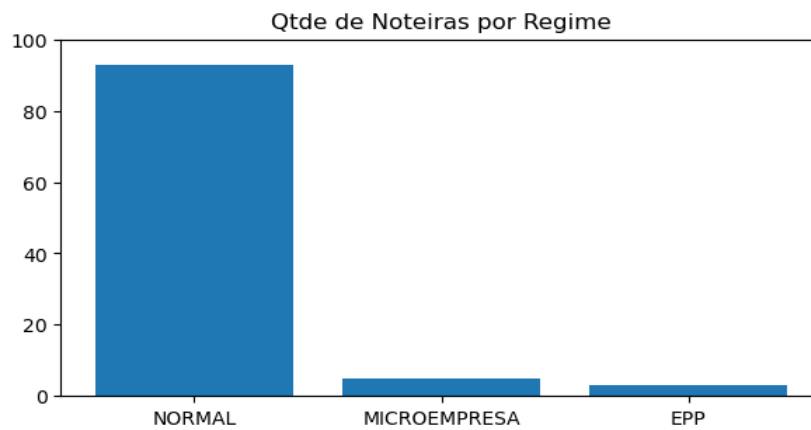
XAVIER, Otávio C. et al. **Identificação de evasão fiscal utilizando dados abertos e inteligência artificial**. RAP (2022), Rio de Janeiro, RJ. P.426-440

## APÊNDICE A – ANÁLISE EXPLORATÓRIA DOS DADOS

### DISTRIBUIÇÃO DAS EMPRESAS NOTEIRAS E NÃO NOTEIRAS POR REGIME

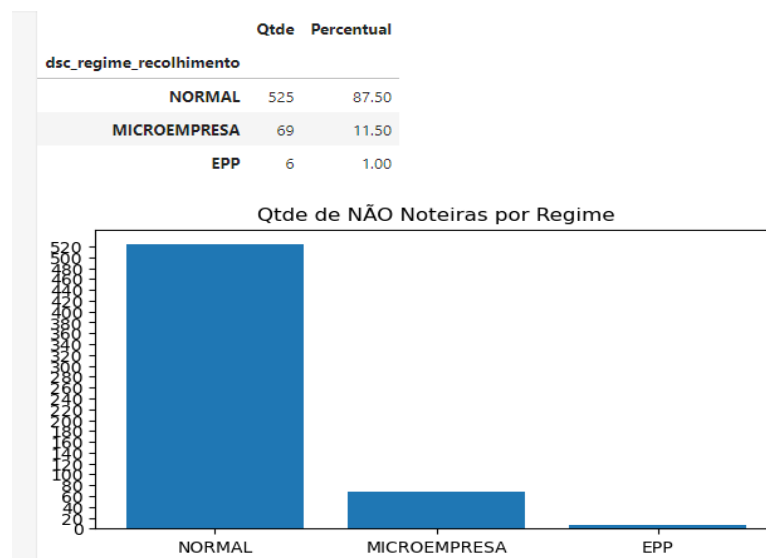
#### - EMPRESAS NOTEIRAS

|                                | Qtde | Percentual |
|--------------------------------|------|------------|
| <b>dsc_regime_recolhimento</b> |      |            |
| <b>NORMAL</b>                  | 93   | 92,08      |
| <b>MICROEMPRESA</b>            | 5    | 4,95       |
| <b>EPP</b>                     | 3    | 2,97       |



Fonte: Elaborado pelo Autor

#### -EMPRESAS NÃO NOTEIRAS

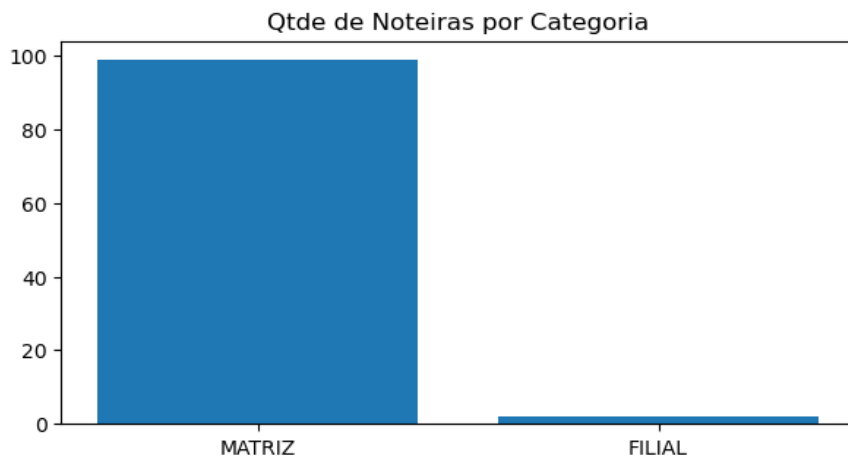


Fonte: Elaborado pelo Autor

## DISTRIBUIÇÃO DAS EMPRESAS NOTEIRAS E NÃO NOTEIRAS POR CATEGORIA

*-EMPRESAS NOTEIRAS*

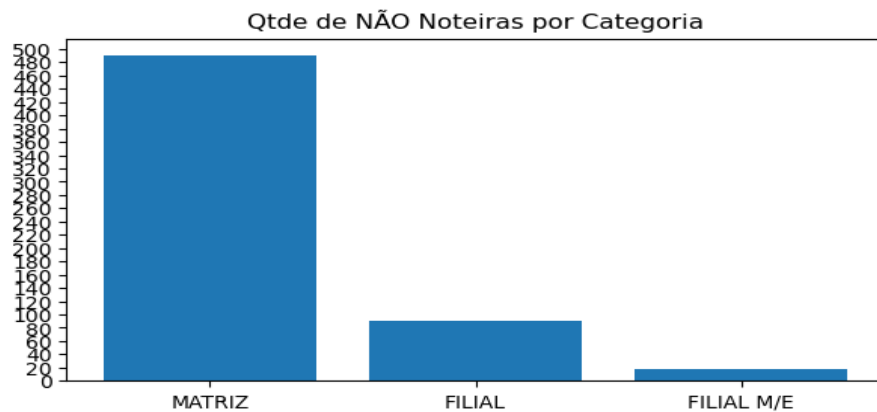
| Qtde                          |    |
|-------------------------------|----|
| dsc_categoria_estabelecimento |    |
| MATRIZ                        | 99 |
| FILIAL                        | 2  |



Fonte: Elaborado pelo Autor

*- EMPRESAS NÃO NOTEIRAS*

| Qtde                          |     |
|-------------------------------|-----|
| dsc_categoria_estabelecimento |     |
| MATRIZ                        | 491 |
| FILIAL                        | 91  |
| FILIAL M/E                    | 18  |

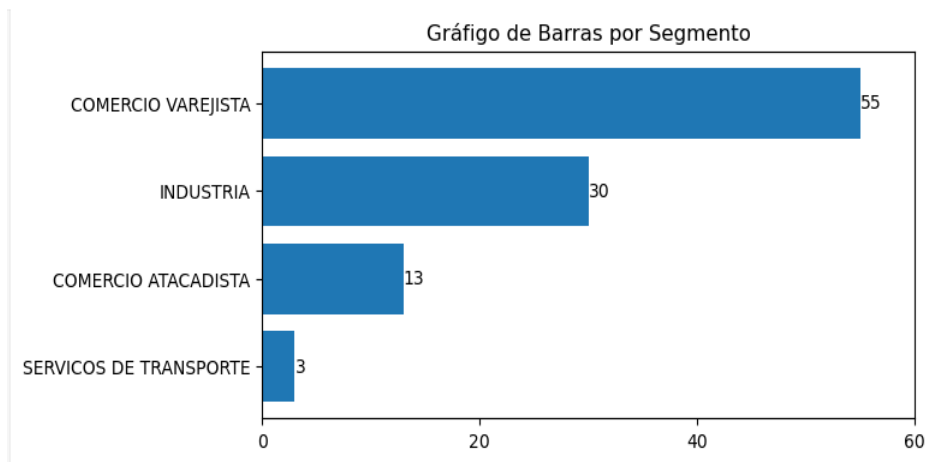


Fonte: Elaborado pelo Autor

## DISTRIBUIÇÃO DAS EMPRESAS NOTEIRAS E NÃO NOTEIRAS POR SEGMENTO ECONÔMICO

### -EMPRESAS NOTEIRAS

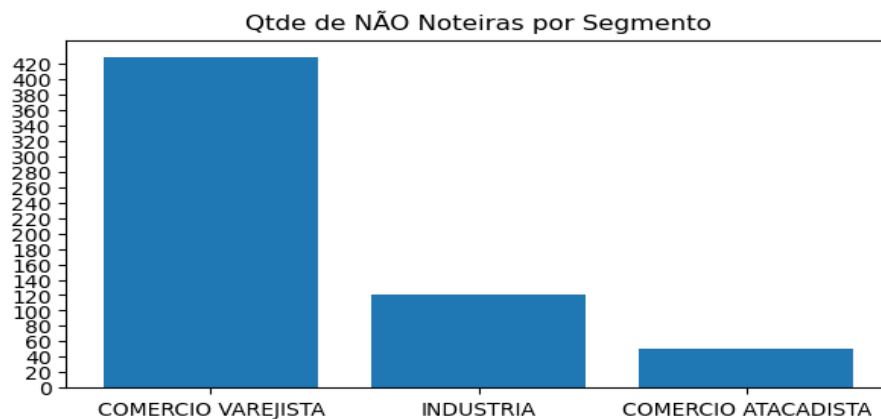
| dsc_segmento_economico | Qtde |
|------------------------|------|
| COMERCIO VAREJISTA     | 55   |
| INDUSTRIA              | 30   |
| COMERCIO ATACADISTA    | 13   |
| SERVICOS DE TRANSPORTE | 3    |



Fonte: Elaborado pelo Autor

### -EMPRESAS NÃO NOTEIRAS

| dsc_segmento_economico | Qtde |
|------------------------|------|
| COMERCIO VAREJISTA     | 429  |
| INDUSTRIA              | 120  |
| COMERCIO ATACADISTA    | 51   |



Fonte: Elaborado pelo Autor

## DISTRIBUIÇÃO DAS EMPRESAS NOTEIRAS POR CNAE

| <b>cod_cnae_primario</b> | <b>dsc_cnae_primario</b>                                     | <b>Qtde</b> |
|--------------------------|--|-------------|
| 4712100                  | Comércio varejista de mercadorias em geral, com predominânci | 34          |
| 1066000                  | Fabricação de alimentos para animais                         | 9           |
| 2441501                  | Produção de alumínio e suas ligas em formas primárias        | 6           |
| 4729699                  | Comércio varejista de produtos alimentícios em geral ou espe | 6           |
| 4644301                  | Comércio atacadista de medicamentos e drogas de uso humano   | 4           |
| 4632001                  | Comércio atacadista de cereais e leguminosas beneficiados    | 3           |
| 4930202                  | Transporte rodoviário de carga, exceto produtos perigosos e  | 3           |
| 4772500                  | Comércio varejista de cosméticos, produtos de perfumaria e d | 2           |
| 1099602                  | Fabricação de pós alimentícios                               | 2           |
| 4744099                  | Comércio varejista de materiais de construção em geral       | 2           |

Fonte: Elaborado pelo Autor

## DISTRIBUIÇÃO DAS EMPRESAS NOTEIRAS POR LOCALIZAÇÃO

| <b>CONTAGEM DE NOTEIRAS POR LOCALIZAÇÃO</b> |                   |                    |                        |
|---|-------------------|--------------------|------------------------|
| <b>Localização</b>                          | <b>Quantidade</b> | <b>Porcentagem</b> | <b>Porc. Acumulada</b> |
| FORTALEZA (capital)                         | 32                | 31.68%             | 31.68%                 |
| INTERIOR                                    | 69                | 68.32%             | 100%                   |

Fonte: Elaborado pelo Autor

## DISTRIBUIÇÃO DAS EMPRESAS NOTEIRAS PELO ANO DE INÍCIO DE ATIVIDADE

| <b>Ano</b>        | <b>Quantidade</b> |
|-------------------|-------------------|
| 2023              | 9                 |
| 2022              | 10                |
| 2021              | 8                 |
| 2020              | 13                |
| 2019              | 13                |
| Anteriores à 2019 | 48                |

Fonte: Elaborado pelo Autor

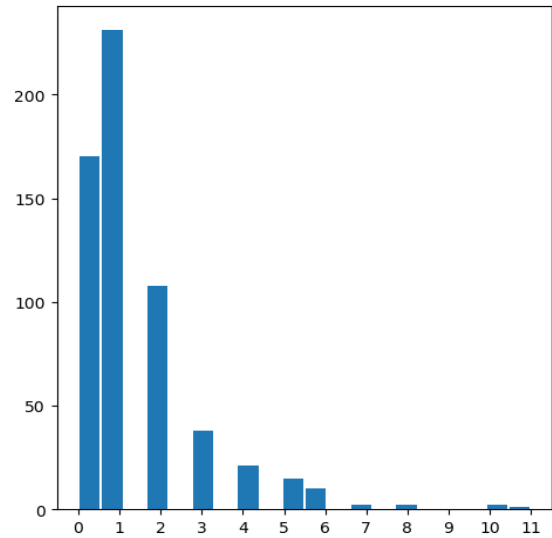
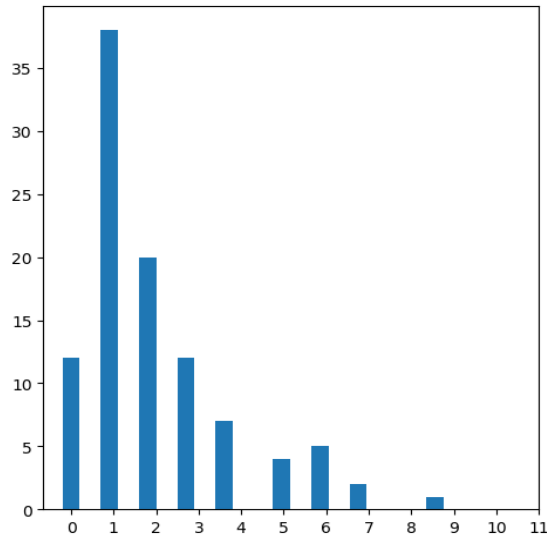


DISTRIBUIÇÃO DAS NOTEIRAS POR TIPO DE CONTADOR (PESSOA/EMPRESA)

| Tipo de Contador | Quantidade |
|------------------|------------|
| Com Contador     | 48         |
| Sem Contador     | 53         |

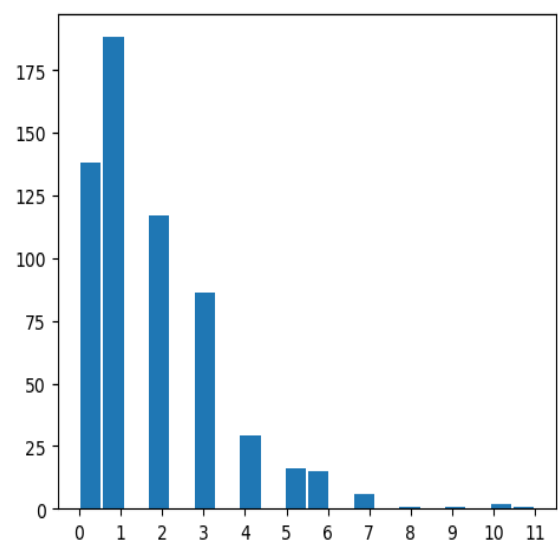
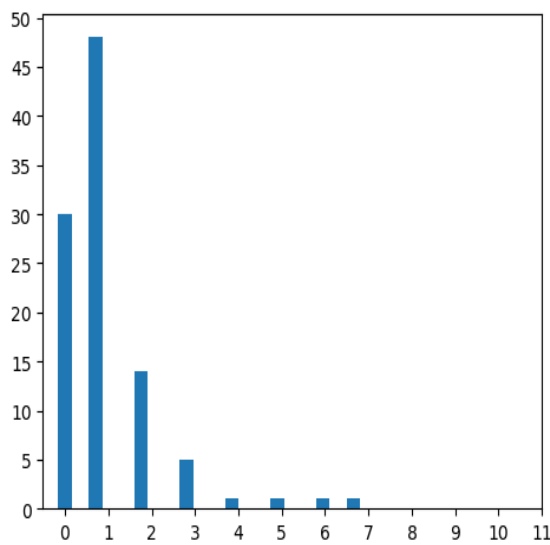
Fonte: Elaborado pelo Autor

QUANTIDADE DE MUDANÇAS DE CONTADOR NOTEIRAS X NÃO NOTEIRAS (RESPECTIVAMENTE)



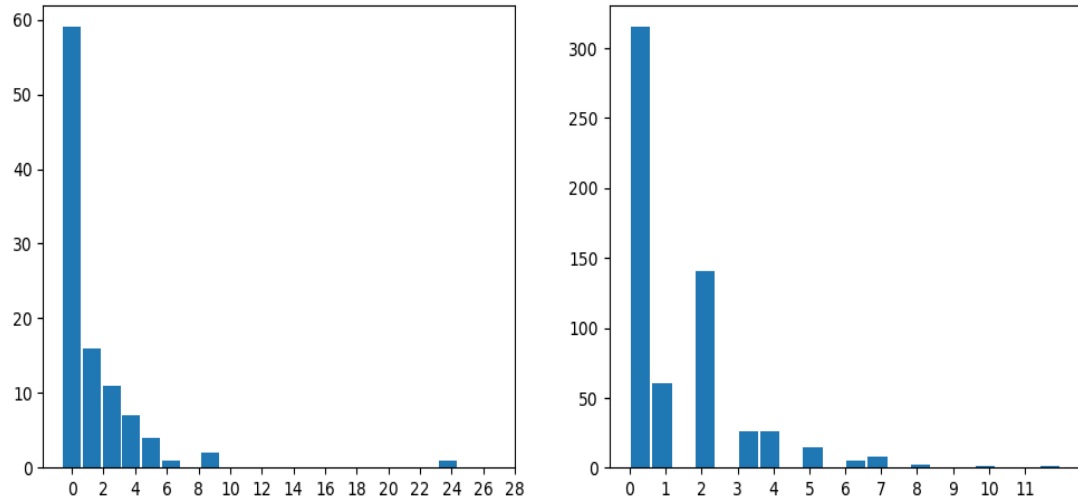
Fonte: Elaborado pelo Autor

QUANTIDADE DE MUDANÇAS DE REGIME NOTEIRAS X NÃO NOTEIRAS (RESPECTIVAMENTE)



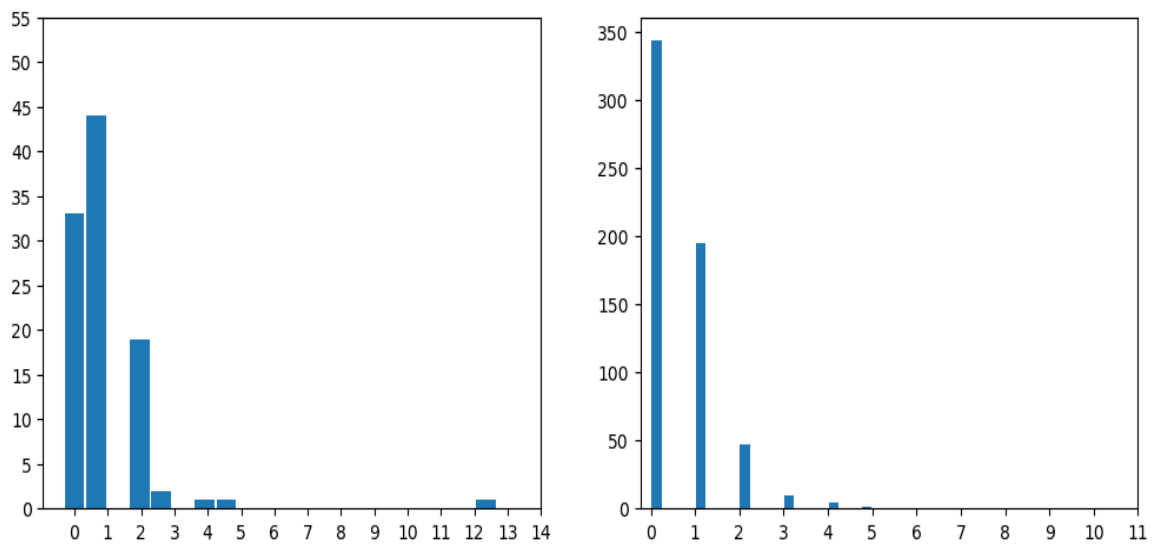
Fonte: Elaborado pelo Autor

QUANTIDADE DE MUDANÇAS DE SITUAÇÃO NOTEIRAS X NÃO NOTEIRAS (RESPECTIVAMENTE)



Fonte: Elaborado pelo Autor

QUANTIDADE DE VEZES EM EDITAL NOTEIRAS X NÃO NOTEIRAS (RESPECTIVAMENTE)



Fonte: Elaborado pelo Autor

## MÉTRICAS ESTATÍSTICA DA ARRECADAÇÃO DAS NOTEIRAS E NÃO NOTEIRAS

*-EMPRESAS NOTEIRAS*

|              | <b>vlr_arrec_1015</b> | <b>vlr_arrec_1023</b> | <b>vlr_arrec_1031</b> | <b>total_arrecadado</b> |
|--------------|-----------------------|-----------------------|-----------------------|-------------------------|
| <b>count</b> | 101                   | 101                   | 101                   | 101                     |
| <b>mean</b>  | R\$ 18.645,07         | R\$ 12.157,30         | R\$ 75.273,41         | R\$ 131.514,75          |
| <b>std</b>   | R\$ 115.633,65        | R\$ 58.718,13         | R\$ 337.657,30        | R\$ 446.229,46          |
| <b>min</b>   | R\$ 0,00              | R\$ 0,00              | R\$ 0,00              | R\$ 0,00                |
| <b>10%</b>   | R\$ 0,00              | R\$ 0,00              | R\$ 0,00              | R\$ 0,00                |
| <b>25%</b>   | R\$ 0,00              | R\$ 0,00              | R\$ 0,00              | R\$ 85,66               |
| <b>50%</b>   | R\$ 0,00              | R\$ 0,00              | R\$ 0,00              | R\$ 8.552,21            |
| <b>75%</b>   | R\$ 0,00              | R\$ 85,66             | R\$ 10.823,94         | R\$ 55.884,28           |
| <b>90%</b>   | R\$ 16.199,56         | R\$ 7.754,89          | R\$ 88.634,09         | R\$ 247.429,92          |
| <b>95%</b>   | R\$ 32.652,66         | R\$ 27.391,55         | R\$ 230.332,22        | R\$ 473.865,49          |
| <b>max</b>   | R\$ 1.117.085,04      | R\$ 498.642,82        | R\$ 2.599.803,78      | R\$ 3.392.702,35        |

Fonte: Elaborado pelo Autor

*- EMPRESAS NÃO NOTEIRAS*

|              | <b>vlr_arrec_1015</b> | <b>vlr_arrec_1023</b> | <b>vlr_arrec_1031</b> | <b>total_arrecadado</b> |
|--------------|-----------------------|-----------------------|-----------------------|-------------------------|
| <b>count</b> | 600                   | 600                   | 600                   | 600                     |
| <b>mean</b>  | R\$ 1.137.340,49      | R\$ 92.307,42         | R\$ 205.841,90        | R\$ 2.157.001,02        |
| <b>std</b>   | R\$ 18.449.222,55     | R\$ 1.146.287,92      | R\$ 1.627.053,67      | R\$ 30.431.984,61       |
| <b>min</b>   | R\$ 0,00              | R\$ 0,00              | R\$ 0,00              | R\$ 0,00                |
| <b>10%</b>   | R\$ 0,00              | R\$ 0,00              | R\$ 0,00              | R\$ 0,00                |
| <b>25%</b>   | R\$ 0,00              | R\$ 0,00              | R\$ 0,00              | R\$ 0,00                |
| <b>50%</b>   | R\$ 0,00              | R\$ 0,00              | R\$ 0,00              | R\$ 2.415,15            |
| <b>75%</b>   | R\$ 0,00              | R\$ 373,20            | R\$ 1.866,38          | R\$ 27.820,25           |
| <b>90%</b>   | R\$ 30.689,60         | R\$ 8.715,50          | R\$ 55.145,37         | R\$ 266.574,51          |
| <b>95%</b>   | R\$ 181.375,88        | R\$ 50.188,70         | R\$ 288.473,86        | R\$ 1.366.299,16        |
| <b>max</b>   | R\$ 379.414.919,85    | R\$ 21.433.178,21     | R\$ 26.683.632,58     | R\$ 583.424.326,51      |

Fonte: Elaborado pelo Autor

## TABELA DE FREQUÊNCIA DE ARRECADAÇÃO DAS NOTEIRAS E NÃO NOTEIRAS

*-EMPRESAS NOTEIRAS*

| <b>faixa_arrec</b>         | <b>Quantidade</b> | <b>Frequência</b> | <b>Frequência Acumul.</b> |
|----------------------------|-------------------|-------------------|---------------------------|
| <b>0-50.000</b>            | 74                | 73.27             | 73.27                     |
| <b>50.001-100.0000</b>     | 9                 | 8.91              | 82.18                     |
| <b>100.001-200.000</b>     | 6                 | 5.94              | 88.12                     |
| <b>200.001-500.000</b>     | 7                 | 6.93              | 95.05                     |
| <b>500.001-1.000.000</b>   | 1                 | 0.99              | 96.04                     |
| <b>1.000.001-5.000.000</b> | 4                 | 3.96              | 100.00                    |

Fonte: Elaborado pelo Autor

*- EMPRESAS NÃO NOTEIRAS*

| <b>faixa_arrec</b>         | <b>Quantidade</b> | <b>Frequência</b> | <b>Frequência Acumul.</b> |
|----------------------------|-------------------|-------------------|---------------------------|
| <b>0-50.000</b>            | 476               | 81.09             | 81.09                     |
| <b>50.001-100.0000</b>     | 35                | 5.96              | 87.05                     |
| <b>100.001-200.000</b>     | 20                | 3.41              | 90.46                     |
| <b>200.001-500.000</b>     | 20                | 3.41              | 93.87                     |
| <b>500.001-1.000.000</b>   | 11                | 1.87              | 95.74                     |
| <b>1.000.001-5.000.000</b> | 25                | 4.26              | 100.00                    |

Fonte: Elaborado pelo Autor

## MÉTRICAS ESTATÍSTICA DOS DÉBITOS DAS NOTEIRAS E NÃO NOTEIRAS

*-EMPRESAS NÃO NOTEIRAS*

|              | <b>vlr_debito_1015</b> | <b>vlr_debito_1023</b> | <b>vlr_debito_1031</b> | <b>vlr_total_debito</b> |
|--------------|------------------------|------------------------|------------------------|-------------------------|
| <b>count</b> | 101                    | 101                    | 101                    | 101                     |
| <b>mean</b>  | R\$ 621,07             | R\$ 0,00               | R\$ 156,44             | R\$ 33.374,65           |
| <b>std</b>   | R\$ 6.241,72           | R\$ 0,00               | R\$ 1.571,96           | R\$ 296.474,19          |
| <b>min</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>10%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>25%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>50%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>75%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>90%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>95%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>max</b>   | R\$ 62.728,48          | R\$ 0,00               | R\$ 15.798,06          | R\$ 2.955.295,39        |

Fonte: Elaborado pelo Autor

*-EMPRESAS NÃO NOTEIRAS*

|              | <b>vlr_debito_1015</b> | <b>vlr_debito_1023</b> | <b>vlr_debito_1031</b> | <b>vlr_total_debito</b> |
|--------------|------------------------|------------------------|------------------------|-------------------------|
| <b>count</b> | 600                    | 600                    | 600                    | 600                     |
| <b>mean</b>  | R\$ 103,49             | R\$ 92,53              | R\$ 84,22              | R\$ 4.067,39            |
| <b>std</b>   | R\$ 2.010,54           | R\$ 1.946,74           | R\$ 1.596,68           | R\$ 42.260,12           |
| <b>min</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>10%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>25%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>50%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>75%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>90%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>95%</b>   | R\$ 0,00               | R\$ 0,00               | R\$ 0,00               | R\$ 0,00                |
| <b>max</b>   | R\$ 47.716,60          | R\$ 46.969,58          | R\$ 38.274,23          | R\$ 691.238,68          |

Fonte: Elaborado pelo Autor

DISTRIBUIÇÃO DE EMPRESAS NOTEIRAS E NÃO NOTEIRAS POR FAIXA DE EMISSÃO DE VALORES DA NOTA FISCAL

*-EMPRESAS NOTEIRAS*

| <b>faixa_emissao</b>             | <b>Quantidade</b> | <b>Frequência</b> | <b>Frequência Acumul.</b> |
|----------------------------------|-------------------|-------------------|---------------------------|
| <b>0-1.000.0000</b>              | 45                | 44.55             | 44.55                     |
| <b>1.000.001-25.000.000</b>      | 19                | 18.81             | 63.37                     |
| <b>25.000.001-50.000.000</b>     | 20                | 19.80             | 83.17                     |
| <b>50.000.001-100.000.000</b>    | 9                 | 8.91              | 92.08                     |
| <b>100.000.001-500.000.000</b>   | 7                 | 6.93              | 99.01                     |
| <b>500.000.000-2.000.000.000</b> | 1                 | 0.99              | 100.00                    |

Fonte: Elaborado pelo Autor

*- EMPRESAS NÃO NOTEIRAS*

| <b>faixa_emissao</b>             | <b>Quantidade</b> | <b>Frequência</b> | <b>Frequência Acumul.</b> |
|----------------------------------|-------------------|-------------------|---------------------------|
| <b>0-1.000.0000</b>              | 498               | 83.28             | 83.28                     |
| <b>1.000.001-25.000.000</b>      | 46                | 7.69              | 90.97                     |
| <b>25.000.001-50.000.000</b>     | 21                | 3.51              | 94.48                     |
| <b>50.000.001-100.000.000</b>    | 19                | 3.18              | 97.66                     |
| <b>100.000.001-500.000.000</b>   | 12                | 2.01              | 99.67                     |
| <b>500.000.000-2.000.000.000</b> | 2                 | 0.33              | 100.00                    |

Fonte: Elaborado pelo Autor

## APÊNDICE B – LISTAS DE FEATURES

### FEATURES DO TIPO QUANTITATIVAS MONETÁRIAS

| CATEGORIA      | NOME                          | DESCRIÇÃO                                       |
|----------------|-------------------------------|---|
| Capital Social | vlr_capital_social            | Valor do capital social da empresa              |
| Capital Social | total_capital_social_contador | Total do capital social registrado por contador |
| Débitos        | total_debito_contador         | Total de débitos registrados por contador       |
| Débitos        | vlr_debito_1066               | Valor do débito com o código de receita 1066    |
| Débitos        | vlr_debito_1040               | Valor do débito com o código de receita 1040    |
| Débitos        | vlr_debito_1058               | Valor do débito com o código de receita 1058    |
| Débitos        | vlr_debito_1031               | Valor do débito com o código de receita 1031    |
| Débitos        | vlr_debito_1090               | Valor do débito com o código de receita 1090    |
| Débitos        | vlr_debito_1015               | Valor do débito com o código de receita 1015    |
| Débitos        | vlr_debito_1023               | Valor do débito com o código de receita 1023    |
| Débitos        | vlr_debito_2020               | Valor do débito com o código de receita 2020    |
| Débitos        | vlr_debito_1104               | Valor do débito com o código de receita 1104    |
| Débitos        | vlr_debito_1120               | Valor do débito com o código de receita 1120    |
| Débitos        | vlr_debito_outras             | Valor do débito com outros códigos de receita   |
| Débitos        | vlr_total_debito              | Valor total de débitos                          |
| Arrecadação    | vlr_arrec_1015                | Valor arrecadado com o código de receita 1015   |
| Arrecadação    | vlr_arrec_1058                | Valor arrecadado com o código de receita 1058   |
| Arrecadação    | vlr_arrec_1031                | Valor arrecadado com o código de receita 1031   |
| Arrecadação    | vlr_arrec_1082                | Valor arrecadado com o código de receita 1082   |
| Arrecadação    | vlr_arrec_1023                | Valor arrecadado com o código de receita 1023   |
| Arrecadação    | vlr_arrec_2020                | Valor arrecadado com o código de receita 2020   |
| Arrecadação    | vlr_arrec_1104                | Valor arrecadado com o código de receita 1104   |

|                         |                                      |  |
|-------------------------|--------------------------------------|--|
| Arrecadação             | vlr_arrec_1201                       | Valor arrecadado com o código de receita 1201  |
| Arrecadação             | vlr_arrec_1155                       | Valor arrecadado com o código de receita 1155  |
| Arrecadação             | vlr_arrec_1090                       | Valor arrecadado com o código de receita 1090  |
| Arrecadação             | vlr_arrec_outras                     | Valor arrecadado com outros códigos de receita   |
| Arrecadação             | total_arrecadado                     | Valor total arrecadado   |
| Notas Fiscais Emitidas  | vlr_tot_nf_icms_emitida              | Valor total de ICMS das notas fiscais emitidas   |
| Notas Fiscais Emitidas  | vlr_q1_nf_icms_emitida               | Primeiro quartil do valor de ICMS das notas fiscais emitidas                                 |
| Notas Fiscais Emitidas  | vlr_q2_nf_icms_emitida               | Segundo quartil do valor de ICMS das notas fiscais emitidas                                  |
| Notas Fiscais Emitidas  | vlr_q3_nf_icms_emitida               | Terceiro quartil do valor de ICMS das notas fiscais emitidas                                 |
| Notas Fiscais Emitidas  | vlr_tot_bc_icms_emitida              | Valor total da base de cálculo de ICMS das notas fiscais emitidas                            |
| Notas Fiscais Emitidas  | vlr_tot_icms_emitida                 | Valor total de ICMS das notas fiscais emitidas   |
| Notas Fiscais Emitidas  | vlr_tot_bc_st_icms_emitida           | Valor total da base de cálculo de substituição tributária de ICMS das notas fiscais emitidas |
| Notas Fiscais Emitidas  | vlr_media_nf_icms_emitida            | Valor médio das notas fiscais de ICMS emitidas   |
| Notas Fiscais Emitidas  | vlr_tot_ano_nf_icms_emitida          | Valor total de ICMS das notas fiscais emitidas no ano  |
| Notas Fiscais Emitidas  | vlr_tot_12_meses_ant_nf_icms_emitida | Valor total de ICMS das notas fiscais emitidas nos 12 meses anteriores                       |
| Notas Fiscais Emitidas  | vlr_tot_24_meses_ant_nf_icms_emitida | Valor total de ICMS das notas fiscais emitidas nos 24 meses anteriores                       |
| Notas Fiscais Emitidas  | vlr_tot_36_meses_ant_nf_icms_emitida | Valor total de ICMS das notas fiscais emitidas nos 36 meses anteriores                       |
| Notas Fiscais Emitidas  | vlr_total_nfe_interestadual_emitida  | Valor total das notas fiscais eletrônicas emitidas nas operações interestaduais              |
| Notas Fiscais Emitidas  | vlr_total_nfe_interna_emitida        | Valor total das notas fiscais eletrônicas emitidas nas operações internas                    |
| Notas Fiscais Recebidas | vlr_tot_nf_icms_recebida             | Valor total de ICMS das notas fiscais recebidas  |
| Notas Fiscais Recebidas | vlr_q1_nf_icms_recebida              | Primeiro quartil do valor de ICMS das notas fiscais recebidas                                |
| Notas Fiscais Recebidas | vlr_q2_nf_icms_recebida              | Segundo quartil do valor de ICMS das notas fiscais recebidas                                 |
| Notas Fiscais Recebidas | vlr_q3_nf_icms_recebida              | Terceiro quartil do valor de ICMS das notas fiscais recebidas                                |
| Notas Fiscais Recebidas | vlr_tot_bc_icms_recebida             | Valor total da base de cálculo de ICMS das notas fiscais recebidas                           |
| Notas Fiscais Recebidas | vlr_tot_icms_recebida                | Valor total de ICMS das notas fiscais recebidas  |



|                         |                                       |   |
|-------------------------|---------------------------------------|---|
| Notas Fiscais Recebidas | vlr_tot_bc_st_icms_recebida           | Valor total da base de cálculo de substituição tributária de ICMS das notas fiscais recebidas |
| Notas Fiscais Recebidas | vlr_media_nf_icms_recebida            | Valor médio das notas fiscais de ICMS recebidas   |
| Notas Fiscais Recebidas | vlr_tot_ano_nf_icms_recebida          | Valor total de ICMS das notas fiscais recebidas no ano  |
| Notas Fiscais Recebidas | vlr_tot_12_meses_ant_nf_icms_recebida | Valor total de ICMS das notas fiscais recebidas nos 12 meses anteriores                       |
| Notas Fiscais Recebidas | vlr_tot_24_meses_ant_nf_icms_recebida | Valor total de ICMS das notas fiscais recebidas nos 24 meses anteriores                       |
| Notas Fiscais Recebidas | vlr_tot_36_meses_ant_nf_icms_recebida | Valor total de ICMS das notas fiscais recebidas nos 36 meses anteriores                       |
| Notas Fiscais Recebidas | vlr_total_nfe_interestadual_recebida  | Valor total das notas fiscais eletrônicas recebidas nas operações interestaduais              |
| Notas Fiscais Recebidas | vlr_total_nfe_interna_recebida        | Valor total das notas fiscais eletrônicas recebidas nas operações internas                    |

Fonte: Elaborado pelo Autor

## FEATURES DO TIPO QUANTITATIVAS NÃO MONETÁRIAS

| CATEGORIA                    | NOME                                     | DESCRIÇÃO  |
|------------------------------|--|--|
| Valores Acumulados           | vlr_part_acum_top10_dest                 | Valor acumulado dos 10 principais destinatários                                    |
| Valores Acumulados           | vlr_part_acum_top10_fornecedor           | Valor acumulado dos 10 principais fornecedores                                     |
| Percentuais de Notas Fiscais | vlr_perc_nfe_valor_1_a_5k_emitida        | Percentual de notas fiscais emitidas com valor entre R\$ 1.000,00 e R\$ 5.000,00   |
| Percentuais de Notas Fiscais | vlr_perc_nfe_valor_5k_a_10k_emitida      | Percentual de notas fiscais emitidas com valor entre R\$ 5.000,00 e R\$ 10.000,00  |
| Percentuais de Notas Fiscais | vlr_perc_nfe_valor_10k_a_50k_emitida     | Percentual de notas fiscais emitidas com valor entre R\$ 10.000,00 e R\$ 50.000,00 |
| Percentuais de Notas Fiscais | vlr_perc_nfe_valor_maior_que_50k_emitida | Percentual de notas fiscais  |

|                              |   |   |
|------------------------------|---|---|
|                              |   | emitidas com valor superior a R\$ 50.000,00   |
| Percentuais de Notas Fiscais | vlr_perc_nfe_valor_1_a_5k_recebida        | Percentual de notas fiscais recebidas com valor entre R\$ 1.000,00 e R\$ 5.000,00   |
| Percentuais de Notas Fiscais | vlr_perc_nfe_valor_5k_a_10k_recebida      | Percentual de notas fiscais recebidas com valor entre R\$ 5.000,00 e R\$ 10.000,00  |
| Percentuais de Notas Fiscais | vlr_perc_nfe_valor_10k_a_50k_recebida     | Percentual de notas fiscais recebidas com valor entre R\$ 10.000,00 e R\$ 50.000,00 |
| Percentuais de Notas Fiscais | vlr_perc_nfe_valor_maior_que_50k_recebida | Percentual de notas fiscais recebidas com valor superior a R\$ 50.000,00            |
| Quantidades de CGFs e CNPJs  | qtde_cgfs_cnpj_base                       | Quantidade de CGFs da base CNPJ   |
| Atividades e Mudanças        | qtde_dias_atividade                       | Quantidade de dias de atividade   |
| Atividades e Mudanças        | qtde_mudancas_cadastro                    | Quantidade de mudanças no cadastro  |
| Atividades e Mudanças        | qtde_baixas                               | Quantidade de baixas  |
| Atividades e Mudanças        | qtde_ativo_edital                         | Quantidade de ativos em edital  |
| Atividades e Mudanças        | qtde_mudancas_contador                    | Quantidade de mudanças de contador  |
| Atividades e Mudanças        | qtde_mudancas_situacao                    | Quantidade de mudanças de situação  |
| Atividades e Mudanças        | qtde_mudancas_regime                      | Quantidade de mudanças de regime  |
| Quantidades de CGFs e CNPJs  | qtde_cgfs_com_debito_contador             | Quantidade de CGFs com débito do contador   |
| Quantidades de CGFs e CNPJs  | qtde_cgfs_contador                        | Quantidade de CGFs do contador  |

|                                       |                                     |   |
|---------------------------------------|-------------------------------------|---|
| Quantidades de CGFs e CNPJs           | qtde_cnpj_base_contador             | Quantidade de CNPJs base do contador                                |
| Percentuais de Atividades do Contador | perc_baixados_contador              | Percentual de baixados do contador                                  |
| Percentuais de Atividades do Contador | perc_ativo_em_edital_contador       | Percentual de ativos em edital do contador                          |
| Percentuais de Atividades do Contador | perc_com_debito_contador            | Percentual de CGFs com débito do contador                           |
| Quantidades de CGFs e CNPJs           | qtde_cgfs_normal_contador           | Quantidade de CGFs normal do contador                               |
| Quantidades de CGFs e CNPJs           | qtde_cgfs_simples_contador          | Quantidade de CGFs do Simples Nacional do contador                  |
| Quantidades de CGFs e CNPJs           | qtde_cgfs_st_contador               | Quantidade de CGFs do regime de Substituição Tributária do contador |
| Quantidades de CGFs e CNPJs           | qtde_cgfs_mei_contador              | Quantidade de CGFs do MEI do contador                               |
| Quantidades de CGFs e CNPJs           | qtde_cgfs_ativo_contador            | Quantidade de CGFs ativos do contador                               |
| Quantidades de CGFs e CNPJs           | qtde_cgfs_ativo_em_edital_contador  | Quantidade de CGFs ativos em edital do contador                     |
| Quantidades de CGFs e CNPJs           | qtde_cgfs_baixados_contador         | Quantidade de CGFs baixados do contador                             |
| Quantidades de CGFs e CNPJs           | qtde_cgfs_cancel_anul_excl_contador | Quantidade de CGFs cancelados, anulados ou excluídos do contador    |
| Notas Fiscais Emitidas e Recebidas    | qtde_notas_fiscais_emitida          | Quantidade de notas fiscais emitidas                                |
| Notas Fiscais Emitidas e Recebidas    | qtde_notas_fiscais_recebida         | Quantidade de notas fiscais recebidas                               |
| Destinatários e Fornecedores          | qtde_destinatario_interestadual     | Quantidade de destinatários interestaduais                          |

|                                    |                                 |  |
|------------------------------------|---------------------------------|--|
| Destinatários e Fornecedores       | qtde_nfe_interestadual_emitida  | Quantidade de notas fiscais eletrônicas interestaduais emitidas                              |
| Destinatários e Fornecedores       | qtde_destinatario_nfe_interna   | Quantidade de destinatários de notas fiscais eletrônicas internas                            |
| Destinatários e Fornecedores       | qtde_nfe_interna_emitida        | Quantidade de notas fiscais eletrônicas internas emitidas                                    |
| Destinatários e Fornecedores       | qtde_destinatario_nfe           | Quantidade de destinatários de notas fiscais eletrônicas                                     |
| Destinatários e Fornecedores       | qtde_fornecedor_interestadual   | Quantidade de fornecedores interestaduais  |
| Destinatários e Fornecedores       | qtde_nfe_interestadual_recebida | Quantidade de notas fiscais eletrônicas interestaduais recebidas                             |
| Destinatários e Fornecedores       | qtde_fornecedor_nfe_interna     | Quantidade de fornecedores de notas fiscais eletrônicas internas                             |
| Destinatários e Fornecedores       | qtde_nfe_interna_recebida       | Quantidade de notas fiscais eletrônicas internas recebidas                                   |
| Notas Fiscais Emitidas e Recebidas | qtde_fornecedores_nfe           | Quantidade de fornecedores de notas fiscais eletrônicas                                      |
| Notas Fiscais Emitidas e Recebidas | qtde_nfe_valor_1_a_5k_emitida   | Quantidade de notas fiscais eletrônicas emitidas com valor entre R\$ 1.000,00 e R\$ 5.000,00 |
| Notas Fiscais Emitidas e Recebidas | qtde_nfe_valor_5k_a_10k_emitida | Quantidade de notas fiscais eletrônicas emitidas com valor entre R\$                         |

|                                    |                                       |   |
|------------------------------------|---------------------------------------|---|
|                                    |                                       | 5.000,00 e R\$ 10.000,00  |
| Notas Fiscais Emitidas e Recebidas | qtde_nfe_valor_10k_a_50k_emitida      | Quantidade de notas fiscais eletrônicas emitidas com valor entre R\$ 10.000,00 e R\$ 50.000,00  |
| Notas Fiscais Emitidas e Recebidas | qtde_nfe_valor_maior_que_50k_emitida  | Quantidade de notas fiscais eletrônicas emitidas com valor superior a R\$ 50.000,00             |
| Notas Fiscais Emitidas e Recebidas | qtde_nfe_valor_1_a_5k_recebida        | Quantidade de notas fiscais eletrônicas recebidas com valor entre R\$ 1.000,00 e R\$ 5.000,00   |
| Notas Fiscais Emitidas e Recebidas | qtde_nfe_valor_5k_a_10k_recebida      | Quantidade de notas fiscais eletrônicas recebidas com valor entre R\$ 5.000,00 e R\$ 10.000,00  |
| Notas Fiscais Emitidas e Recebidas | qtde_nfe_valor_10k_a_50k_recebida     | Quantidade de notas fiscais eletrônicas recebidas com valor entre R\$ 10.000,00 e R\$ 50.000,00 |
| Notas Fiscais Emitidas e Recebidas | qtde_nfe_valor_maior_que_50k_recebida | Quantidade de notas fiscais eletrônicas recebidas com valor superior a R\$ 50.000,00            |

Fonte: Elaborado pelo Autor

## FEATURES SELECIONADAS

| <b>NOME</b>                                      | <b>TIPO</b>   |
|--|---------------|
| pipeline-1__vlr_capital_social                   | Monetária     |
| pipeline-1__total_capital_social_contador        | Monetária     |
| pipeline-1__vlr_tot_nf_icms_emitida              | Monetária     |
| pipeline-1__vlr_q1_nf_icms_emitida               | Monetária     |
| pipeline-1__vlr_q2_nf_icms_emitida               | Monetária     |
| pipeline-1__vlr_q3_nf_icms_emitida               | Monetária     |
| pipeline-1__vlr_media_nf_icms_emitida            | Monetária     |
| pipeline-1__vlr_tot_ano_nf_icms_emitida          | Monetária     |
| pipeline-1__vlr_tot_12_meses_ant_nf_icms_emitida | Monetária     |
| pipeline-1__vlr_tot_nf_icms_recebida             | Monetária     |
| pipeline-1__vlr_q1_nf_icms_recebida              | Monetária     |
| pipeline-1__vlr_q2_nf_icms_recebida              | Monetária     |
| pipeline-1__vlr_q3_nf_icms_recebida              | Monetária     |
| pipeline-1__vlr_media_nf_icms_recebida           | Monetária     |
| pipeline-1__vlr_tot_ano_nf_icms_recebida         | Monetária     |
| pipeline-1__vlr_total_nfe_interna_emitida        | Monetária     |
| pipeline-1__vlr_total_nfe_interna_recebida       | Monetária     |
| pipeline-2__qtde_cgfs_cnpj_base                  | Não Monetária |
| pipeline-2__qtde_dias_atividade                  | Não Monetária |
| pipeline-2__qtde_mudancas_cadastro               | Não Monetária |
| pipeline-2__qtde_mudancas_contador               | Não Monetária |
| pipeline-2__qtde_mudancas_regime                 | Não Monetária |
| pipeline-2__qtde_cgfs_contador                   | Não Monetária |
| pipeline-2__qtde_cnpj_base_contador              | Não Monetária |
| pipeline-2__qtde_cgfs_normal_contador            | Não Monetária |
| pipeline-2__qtde_cgfs_simples_contador           | Não Monetária |
| pipeline-2__qtde_cgfs_baixados_contador          | Não Monetária |
| pipeline-2__qtde_nfe_valor_10k_a_50k_emitida     | Não Monetária |
| pipeline-2__qtde_nfe_valor_maior_que_50k_emitida | Não Monetária |
| pipeline-3__cod_regime_recolhimento_1.0          | Não Monetária |

Fonte: Elaborado pelo Autor

## APÊNDICE C – MÉTRICAS DOS MODELOS AVALIADOS

### MÉTRICAS PARA O MODELO RANDOM FOREST - 1 AMOSTRA

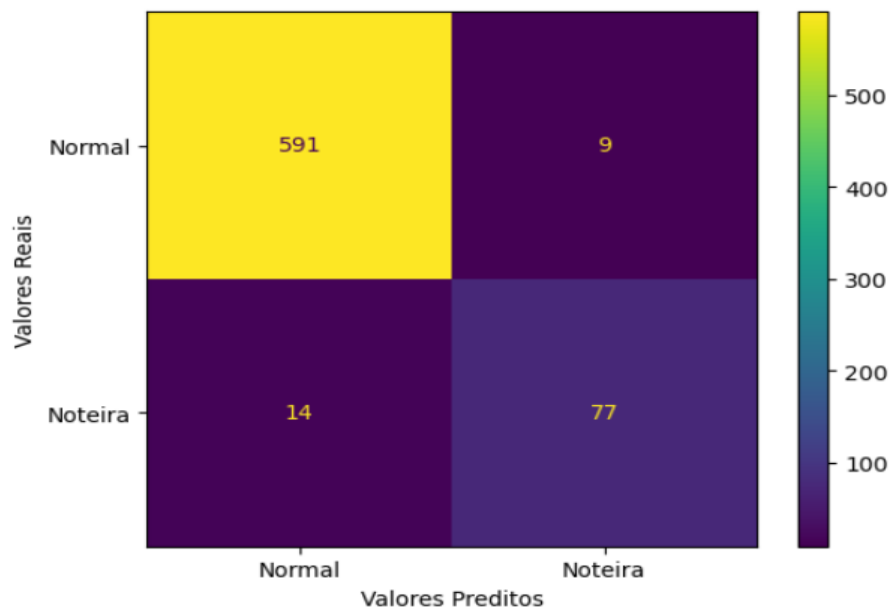
#### Classification Report por CGF no Dataset de Teste

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.98      | 0.98   | 0.98     | 600     |
| 1            | 0.90      | 0.85   | 0.87     | 91      |
| accuracy     |           |        | 0.97     | 691     |
| macro avg    | 0.94      | 0.92   | 0.93     | 691     |
| weighted avg | 0.97      | 0.97   | 0.97     | 691     |

Fonte: Elaborado pelo Autor

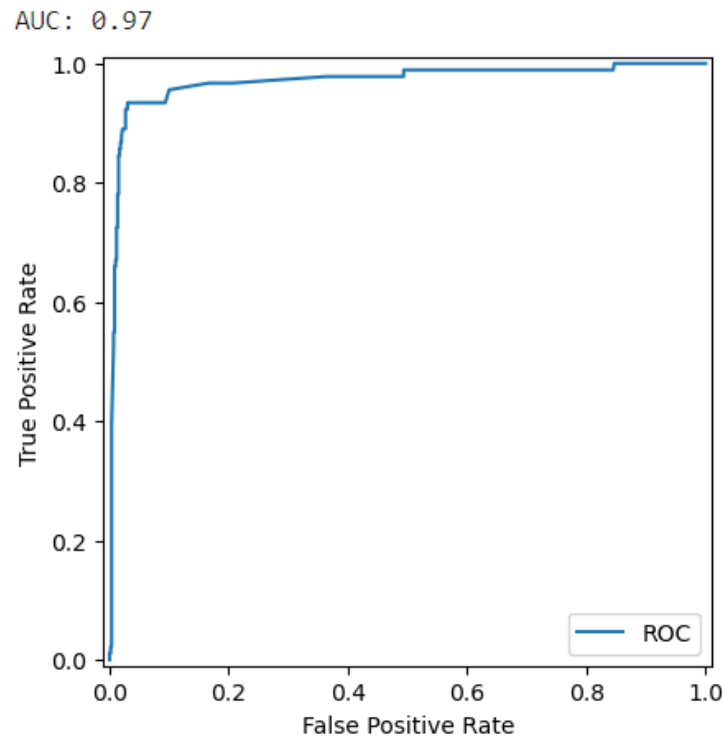
### MATRIZ DE CONFUSÃO PARA O MODELO RANDOM FOREST - 1 AMOSTRA

#### Matriz de Confusão por CGF no Dataset de Teste



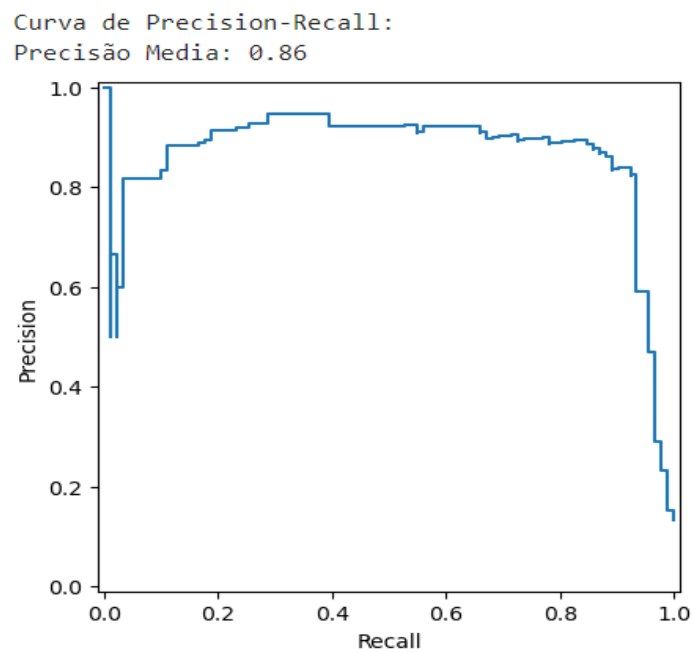
Fonte: Elaborado pelo Autor

## CURVA ROC PARA 1 AMOSTRA - RANDOM FOREST



Fonte: Elaborado pelo Autor

## CURVA PRECISION-RECALL PARA 1 AMOSTRA - RANDOM FOREST



Fonte: Elaborado pelo Autor



## MÉTRICAS PARA TODOS OS MODELOS AVALIADOS PARA 1 AMOSTRA

|                     |        | precision | recall   | f1-score | accuracy | support |
|---------------------|--------|-----------|----------|----------|----------|---------|
| Modelo              | Classe |           |          |          |          |         |
| Regressao_Logistica | 0      | 0.963455  | 0.966667 | 0.965058 | 0.939219 | 600.0   |
|                     | 1      | 0.775281  | 0.758242 | 0.766667 | 0.939219 | 91.0    |
| KNN                 | 0      | 0.939799  | 0.936667 | 0.938230 | 0.892909 | 600.0   |
|                     | 1      | 0.591398  | 0.604396 | 0.597826 | 0.892909 | 91.0    |
| Rede_Neural         | 0      | 0.958541  | 0.963333 | 0.960931 | 0.931983 | 600.0   |
|                     | 1      | 0.750000  | 0.725275 | 0.737430 | 0.931983 | 91.0    |
| Random_Forest       | 0      | 0.976860  | 0.985000 | 0.980913 | 0.966715 | 600.0   |
|                     | 1      | 0.895349  | 0.846154 | 0.870056 | 0.966715 | 91.0    |
| XGboost             | 0      | 0.957237  | 0.970000 | 0.963576 | 0.936324 | 600.0   |
|                     | 1      | 0.783133  | 0.714286 | 0.747126 | 0.936324 | 91.0    |

Fonte: Elaborado pelo Autor

## MÉTRICAS AGREGADAS PARA TODOS OS MODELOS PARA 10 AMOSTRAS

## Regressao\_Logistica

|          |           | min        | max        | mean       | std      |
|----------|-----------|------------|------------|------------|----------|
| 0        | f1-score  | 0.951280   | 0.969345   | 0.963396   | 0.005312 |
|          | precision | 0.942717   | 0.963756   | 0.957548   | 0.006648 |
|          | recall    | 0.960000   | 0.975000   | 0.969333   | 0.005455 |
|          | support   | 600.000000 | 600.000000 | 600.000000 | 0.000000 |
| 1        | f1-score  | 0.654971   | 0.788571   | 0.746542   | 0.039390 |
|          | precision | 0.700000   | 0.821429   | 0.779946   | 0.037645 |
|          | recall    | 0.615385   | 0.758242   | 0.716484   | 0.045692 |
|          | support   | 91.000000  | 91.000000  | 91.000000  | 0.000000 |
| accuracy | support   | 0.914616   | 0.946454   | 0.936035   | 0.009349 |

Fonte: Elaborado pelo Autor

## KNN

|                 |                  | min        | max        | mean       | std      |
|-----------------|------------------|------------|------------|------------|----------|
| 0               | <b>f1-score</b>  | 0.932546   | 0.954128   | 0.944844   | 0.007042 |
|                 | <b>precision</b> | 0.933665   | 0.954925   | 0.944405   | 0.006301 |
|                 | <b>recall</b>    | 0.921667   | 0.960000   | 0.945333   | 0.010535 |
|                 | <b>support</b>   | 600.000000 | 600.000000 | 600.000000 | 0.000000 |
| 1               | <b>f1-score</b>  | 0.569832   | 0.699454   | 0.635299   | 0.040832 |
|                 | <b>precision</b> | 0.552381   | 0.700000   | 0.639039   | 0.049324 |
|                 | <b>recall</b>    | 0.560440   | 0.703297   | 0.632967   | 0.042780 |
|                 | <b>support</b>   | 91.000000  | 91.000000  | 91.000000  | 0.000000 |
| <b>accuracy</b> | <b>support</b>   | 0.884226   | 0.920405   | 0.904197   | 0.011949 |

Fonte: Elaborado pelo Autor

## Rede\_Neural

|                 |                  | min        | max        | mean       | std      |
|-----------------|------------------|------------|------------|------------|----------|
| 0               | <b>f1-score</b>  | 0.944582   | 0.970760   | 0.957454   | 0.008923 |
|                 | <b>precision</b> | 0.937603   | 0.973199   | 0.952998   | 0.010413 |
|                 | <b>recall</b>    | 0.945000   | 0.976667   | 0.962000   | 0.010086 |
|                 | <b>support</b>   | 600.000000 | 600.000000 | 600.000000 | 0.000000 |
| 1               | <b>f1-score</b>  | 0.612717   | 0.810811   | 0.708539   | 0.062121 |
|                 | <b>precision</b> | 0.645161   | 0.825000   | 0.733677   | 0.064126 |
|                 | <b>recall</b>    | 0.582418   | 0.824176   | 0.686813   | 0.070886 |
|                 | <b>support</b>   | 91.000000  | 91.000000  | 91.000000  | 0.000000 |
| <b>accuracy</b> | <b>support</b>   | 0.903039   | 0.949349   | 0.925760   | 0.015572 |

Fonte: Elaborado pelo Autor

## Random\_Forest

|                 |                  | min        | max        | mean       | std      |
|-----------------|------------------|------------|------------|------------|----------|
| 0               | <b>f1-score</b>  | 0.964315   | 0.980944   | 0.971768   | 0.005574 |
|                 | <b>precision</b> | 0.960331   | 0.975288   | 0.968081   | 0.004959 |
|                 | <b>recall</b>    | 0.963333   | 0.986667   | 0.975500   | 0.007456 |
|                 | <b>support</b>   | 600.000000 | 600.000000 | 600.000000 | 0.000000 |
| 1               | <b>f1-score</b>  | 0.757062   | 0.868571   | 0.808572   | 0.036076 |
|                 | <b>precision</b> | 0.763441   | 0.904762   | 0.831035   | 0.046172 |
|                 | <b>recall</b>    | 0.736264   | 0.835165   | 0.787912   | 0.033190 |
|                 | <b>support</b>   | 91.000000  | 91.000000  | 91.000000  | 0.000000 |
| <b>accuracy</b> | <b>support</b>   | 0.937771   | 0.966715   | 0.950796   | 0.009648 |

Fonte: Elaborado pelo Autor

## XGboost

|                 |                  | min        | max        | mean       | std      |
|-----------------|------------------|------------|------------|------------|----------|
| 0               | <b>f1-score</b>  | 0.946667   | 0.970149   | 0.959506   | 0.006157 |
|                 | <b>precision</b> | 0.944444   | 0.965347   | 0.955421   | 0.007273 |
|                 | <b>recall</b>    | 0.946667   | 0.975000   | 0.963667   | 0.008157 |
|                 | <b>support</b>   | 600.000000 | 600.000000 | 600.000000 | 0.000000 |
| 1               | <b>f1-score</b>  | 0.648352   | 0.795455   | 0.723797   | 0.041943 |
|                 | <b>precision</b> | 0.648352   | 0.823529   | 0.747131   | 0.046693 |
|                 | <b>recall</b>    | 0.626374   | 0.769231   | 0.703297   | 0.049957 |
|                 | <b>support</b>   | 91.000000  | 91.000000  | 91.000000  | 0.000000 |
| <b>accuracy</b> | <b>support</b>   | 0.907381   | 0.947902   | 0.929378   | 0.010696 |

Fonte: Elaborado pelo Autor

## MÉTRICAS PARA TODAS AS AMOSTRAS

*- REGRESSÃO LOGÍSTICA*

|   | 0        |           |          |         | 1        |           |          |         | accuracy |
|---|----------|-----------|----------|---------|----------|-----------|----------|---------|----------|
|   | f1-score | precision | recall   | support | f1-score | precision | recall   | support | support  |
| 0 | 0.966088 | 0.958949  | 0.973333 | 600.0   | 0.763006 | 0.804878  | 0.725275 | 91.0    | 0.940666 |
| 1 | 0.966833 | 0.962046  | 0.971667 | 600.0   | 0.772727 | 0.800000  | 0.747253 | 91.0    | 0.942113 |
| 2 | 0.964433 | 0.957307  | 0.971667 | 600.0   | 0.751445 | 0.792683  | 0.714286 | 91.0    | 0.937771 |
| 3 | 0.967742 | 0.960591  | 0.975000 | 600.0   | 0.774566 | 0.817073  | 0.736264 | 91.0    | 0.943560 |
| 4 | 0.960000 | 0.960000  | 0.960000 | 600.0   | 0.736264 | 0.736264  | 0.736264 | 91.0    | 0.930535 |
| 5 | 0.965919 | 0.963516  | 0.968333 | 600.0   | 0.770950 | 0.784091  | 0.758242 | 91.0    | 0.940666 |
| 6 | 0.969345 | 0.963756  | 0.975000 | 600.0   | 0.788571 | 0.821429  | 0.758242 | 91.0    | 0.946454 |
| 7 | 0.951280 | 0.942717  | 0.960000 | 600.0   | 0.654971 | 0.700000  | 0.615385 | 91.0    | 0.914616 |
| 8 | 0.959604 | 0.949429  | 0.970000 | 600.0   | 0.710059 | 0.769231  | 0.659341 | 91.0    | 0.929088 |
| 9 | 0.962717 | 0.957166  | 0.968333 | 600.0   | 0.742857 | 0.773810  | 0.714286 | 91.0    | 0.934877 |

Fonte: Elaborado pelo Autor

*- KNN*

|   | 0        |           |          |         | 1        |           |          |         | accuracy |
|---|----------|-----------|----------|---------|----------|-----------|----------|---------|----------|
|   | f1-score | precision | recall   | support | f1-score | precision | recall   | support | support  |
| 0 | 0.932546 | 0.943686  | 0.921667 | 600.0   | 0.591837 | 0.552381  | 0.637363 | 91.0    | 0.884226 |
| 1 | 0.947368 | 0.949749  | 0.945000 | 600.0   | 0.659459 | 0.648936  | 0.670330 | 91.0    | 0.908828 |
| 2 | 0.938436 | 0.936877  | 0.940000 | 600.0   | 0.588889 | 0.595506  | 0.582418 | 91.0    | 0.892909 |
| 3 | 0.949293 | 0.946932  | 0.951667 | 600.0   | 0.659218 | 0.670455  | 0.648352 | 91.0    | 0.911722 |
| 4 | 0.935993 | 0.933665  | 0.938333 | 600.0   | 0.569832 | 0.579545  | 0.560440 | 91.0    | 0.888567 |
| 5 | 0.954128 | 0.954925  | 0.953333 | 600.0   | 0.699454 | 0.695652  | 0.703297 | 91.0    | 0.920405 |
| 6 | 0.948247 | 0.949833  | 0.946667 | 600.0   | 0.663043 | 0.655914  | 0.670330 | 91.0    | 0.910275 |
| 7 | 0.947718 | 0.943802  | 0.951667 | 600.0   | 0.644068 | 0.662791  | 0.626374 | 91.0    | 0.908828 |
| 8 | 0.951280 | 0.942717  | 0.960000 | 600.0   | 0.654971 | 0.700000  | 0.615385 | 91.0    | 0.914616 |
| 9 | 0.943428 | 0.941860  | 0.945000 | 600.0   | 0.622222 | 0.629213  | 0.615385 | 91.0    | 0.901592 |

Fonte: Elaborado pelo Autor

## - REDE NEURAL

|   | 0        |           |          |         | 1        |           |          |         | accuracy |
|---|----------|-----------|----------|---------|----------|-----------|----------|---------|----------|
|   | f1-score | precision | recall   | support | f1-score | precision | recall   | support | support  |
| 0 | 0.958949 | 0.944984  | 0.973333 | 600.0   | 0.695122 | 0.780822  | 0.626374 | 91.0    | 0.927641 |
| 1 | 0.949545 | 0.942529  | 0.956667 | 600.0   | 0.647399 | 0.682927  | 0.615385 | 91.0    | 0.911722 |
| 2 | 0.946578 | 0.948161  | 0.945000 | 600.0   | 0.652174 | 0.645161  | 0.659341 | 91.0    | 0.907381 |
| 3 | 0.970760 | 0.973199  | 0.968333 | 600.0   | 0.810811 | 0.797872  | 0.824176 | 91.0    | 0.949349 |
| 4 | 0.958541 | 0.953795  | 0.963333 | 600.0   | 0.715909 | 0.741176  | 0.692308 | 91.0    | 0.927641 |
| 5 | 0.965975 | 0.961983  | 0.970000 | 600.0   | 0.768362 | 0.790698  | 0.747253 | 91.0    | 0.940666 |
| 6 | 0.967795 | 0.959083  | 0.976667 | 600.0   | 0.771930 | 0.825000  | 0.725275 | 91.0    | 0.943560 |
| 7 | 0.944582 | 0.937603  | 0.951667 | 600.0   | 0.612717 | 0.646341  | 0.582418 | 91.0    | 0.903039 |
| 8 | 0.954281 | 0.951907  | 0.956667 | 600.0   | 0.692737 | 0.704545  | 0.681319 | 91.0    | 0.920405 |
| 9 | 0.957535 | 0.956739  | 0.958333 | 600.0   | 0.718232 | 0.722222  | 0.714286 | 91.0    | 0.926194 |

Fonte: Elaborado pelo Autor

## - RANDOM FOREST

|   | 0        |           |          |         | 1        |           |          |         | accuracy |
|---|----------|-----------|----------|---------|----------|-----------|----------|---------|----------|
|   | f1-score | precision | recall   | support | f1-score | precision | recall   | support | support  |
| 0 | 0.974231 | 0.971808  | 0.976667 | 600.0   | 0.826816 | 0.840909  | 0.813187 | 91.0    | 0.955137 |
| 1 | 0.968333 | 0.968333  | 0.968333 | 600.0   | 0.791209 | 0.791209  | 0.791209 | 91.0    | 0.945007 |
| 2 | 0.972705 | 0.965517  | 0.980000 | 600.0   | 0.809249 | 0.853659  | 0.769231 | 91.0    | 0.952243 |
| 3 | 0.980100 | 0.975248  | 0.985000 | 600.0   | 0.863636 | 0.894118  | 0.835165 | 91.0    | 0.965268 |
| 4 | 0.964942 | 0.966555  | 0.963333 | 600.0   | 0.771739 | 0.763441  | 0.780220 | 91.0    | 0.939219 |
| 5 | 0.980944 | 0.975288  | 0.986667 | 600.0   | 0.868571 | 0.904762  | 0.835165 | 91.0    | 0.966715 |
| 6 | 0.970100 | 0.966887  | 0.973333 | 600.0   | 0.797753 | 0.816092  | 0.780220 | 91.0    | 0.947902 |
| 7 | 0.964315 | 0.960331  | 0.968333 | 600.0   | 0.757062 | 0.779070  | 0.736264 | 91.0    | 0.937771 |
| 8 | 0.971761 | 0.968543  | 0.975000 | 600.0   | 0.808989 | 0.827586  | 0.791209 | 91.0    | 0.950796 |
| 9 | 0.970248 | 0.962295  | 0.978333 | 600.0   | 0.790698 | 0.839506  | 0.747253 | 91.0    | 0.947902 |

Fonte: Elaborado pelo Autor

## - XGBOOST

|          | 0        |           |          |         | 1        |           |          |         | accuracy |
|----------|----------|-----------|----------|---------|----------|-----------|----------|---------|----------|
|          | f1-score | precision | recall   | support | f1-score | precision | recall   | support | support  |
| <b>0</b> | 0.960669 | 0.964706  | 0.956667 | 600.0   | 0.748663 | 0.729167  | 0.769231 | 91.0    | 0.931983 |
| <b>1</b> | 0.957886 | 0.949264  | 0.966667 | 600.0   | 0.701754 | 0.750000  | 0.659341 | 91.0    | 0.926194 |
| <b>2</b> | 0.946667 | 0.946667  | 0.946667 | 600.0   | 0.648352 | 0.648352  | 0.648352 | 91.0    | 0.907381 |
| <b>3</b> | 0.970149 | 0.965347  | 0.975000 | 600.0   | 0.795455 | 0.823529  | 0.769231 | 91.0    | 0.947902 |
| <b>4</b> | 0.962717 | 0.957166  | 0.968333 | 600.0   | 0.742857 | 0.773810  | 0.714286 | 91.0    | 0.934877 |
| <b>5</b> | 0.958541 | 0.953795  | 0.963333 | 600.0   | 0.715909 | 0.741176  | 0.692308 | 91.0    | 0.927641 |
| <b>6</b> | 0.960866 | 0.960067  | 0.961667 | 600.0   | 0.740331 | 0.744444  | 0.736264 | 91.0    | 0.931983 |
| <b>7</b> | 0.953795 | 0.944444  | 0.963333 | 600.0   | 0.670588 | 0.721519  | 0.626374 | 91.0    | 0.918958 |
| <b>8</b> | 0.960866 | 0.960067  | 0.961667 | 600.0   | 0.740331 | 0.744444  | 0.736264 | 91.0    | 0.931983 |
| <b>9</b> | 0.962902 | 0.952692  | 0.973333 | 600.0   | 0.733728 | 0.794872  | 0.681319 | 91.0    | 0.934877 |

Fonte: Elaborado pelo Autor