NOTA CONFORME: SISTEMA INTEGRADO PARA CLASSIFICAÇÃO AUTOMATIZADA DE SERVIÇOS EM NFS-E COM MACHINE LEARNING

TARCISIO PARAISO FARIAS

SALVADOR

2025

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA BAHIA

CAMPUS SALVADOR

TARCISIO PARAISO FARIAS

NOTA CONFORME: SISTEMA INTEGRADO PARA CLASSIFICAÇÃO AUTOMATIZADA DE SERVIÇOS EM NFS-E COM MACHINE LEARNING

SALVADOR

2025

TARCISIO PARAISO FARIAS

NOTA CONFORME: SISTEMA INTEGRADO PARA CLASSIFICAÇÃO AUTOMATIZADA DE SERVIÇOS EM NFS-E COM MACHINE LEARNING

Dissertação apresentada ao Instituto Federal de Educação, Ciência e Tecnologia da Bahia, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Produtos, área de concentração em Sistemas e Produtos Computacionais, de Controle e Comunicação, para a obtenção do título de Mestre.

Orientador: Prof. Dr. Thiago Souto Mendes

SALVADOR

2025

FICHA CATALOGRÁFICA ELABORADA PELO SISTEMA DE BIBLIOTECAS DO IFBA, COM OS DADOS FORNECIDOS PELO(A) AUTOR(A)

F224n Farias, Tarcisio Paraiso

Nota conforme: sistema integrado para classificação automatizada de serviços em NFS-E com machine learning / Tarcisio Paraiso Farias; orientador Thiago Souto Mendes -- Salvador, 2025.

112 p.

Dissertação (Programa de Pós-Graduação em Engenharia de Sistemas de Produtos) -- Instituto Federal da Bahia, 2025.

1. Inteligência artificial. 2. Processamento de linguagem natural. 3. Aprendizado de máquina. 4. Nota fiscal de serviços eletrônica. 5. Auditoria fiscal. I. Mendes, Thiago Souto, orient. II. TÍTULO.

CDU 657.243.1

INSTITUTO FEDERAL DA BAHIA

PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE SISTEMAS E PRODUTOS – PPGESP

NOTA CONFORME: SISTEMA INTEGRADO PARA CLASSIFICAÇÃO AUTOMATIZADA DE SERVIÇOS EM NFS-E MACHINE LEARNING

TARCISIO PARAISO FARIAS

Produto(s) Gerado(s): Dissertação e Artigo Aprovado.

Orientador: Prof. Dr. Thiago Souto Mendes (IFBA/PPGESP)

Banca examinadora:

_____

Prof. Dr. Thiago Souto Mendes

Orientador – (IFBA/PPGESP)

_____

Prof. Dr. Leandro José Silva Andrade

Membro Externo (UFBA)

_____

Prof. Dr. Cleber Jorge Lira de Santana

Membro Interno (IFBA/PPGESP)

Trabalho de Conclusão de Curso aprovado pela banca examinadora em 29/08//2025

Documento assinado eletronicamente por **THIAGO SOUTO MENDES**, **Docente Permanente**, em 29/08/2025, às 14:49, conforme decreto nº 8.539/2015.

Documento assinado eletronicamente por **Leandro José Silva Andrade**, **Usuário Externo**, em 29/08/2025, às 17:20, conforme decreto nº 8.539/2015.

Documento assinado eletronicamente por **CLEBER JORGE LIRA DE SANTANA**, **Docente Permanente**, em 11/09/2025, às 11:23, conforme decreto nº 8.539/2015.

A autenticidade do documento pode ser conferida no site http://sei.ifba.edu.br/sei/controlador_externo.php?acao=documento_conferir&acao_origem=documento_conferir&id_orgao_acesso_externo=0 informando o código verificador **4369976** e o código CRC **5D2AB437**.

DEDICATÓRIA

Dedico esta pesquisa às minhas amadas filhas, cuja presença ilumina minha jornada e o amor me inspira diariamente, dando-me forças para seguir em frente.

AGRADECIMENTOS

A Deus, pela dádiva da vida e pela oportunidade de concretizar mais um sonho.

À minha esposa, pelo apoio incondicional, pela paciência e pelo incentivo constante em cada etapa dessa jornada.

Às minhas filhas, que, dentro das limitações da infância, souberam compreender e aceitar esse período de dedicação intensa.

À minha sogra, que generosamente abdicou de suas próprias demandas para oferecer suporte diário à minha esposa, permitindo-me mergulhar na pesquisa com tranquilidade.

Ao meu orientador, Prof. Dr. Thiago Mendes, por sua orientação precisa, paciência e dedicação ao longo de todo o processo.

Aos professores do Programa de Pós-Graduação em Engenharia de Sistemas e Produtos (PPGESP) pelos valiosos conhecimentos compartilhados ao longo da minha trajetória acadêmica.

À banca de qualificação, composta pelos Professores Dr. Cleber Lira (IFBA) e Dr. Leandro Andrade (UFBA), pelas contribuições, sugestões e críticas construtivas que enriqueceram e fortaleceram esta pesquisa.

À Emprel e à Prefeitura do Recife–PE, pelo acolhimento ao projeto e pelas condições essenciais fornecidas para sua realização.

Ao Rafael Sena, que abraçou essa iniciativa com entusiasmo e a acompanhou de perto; por meio dele, estendo minha gratidão a toda a equipe da Secretaria de Finanças, que de alguma forma contribuiu para o desenvolvimento deste trabalho.

Aos Auditores do Tesouro Municipal, cuja sua participação foi fundamental, desde a estruturação até a validação do projeto.

A Rosana, Laurécio e Calábria, cujo apoio nos encaminhamentos iniciais foram essenciais para que o projeto chegasse às mãos corretas, além de estarem sempre à disposição. Por meio deles, estendo minha gratidão à equipe do Departamento de Inteligência de Dados.

RESUMO

Este trabalho apresenta o desenvolvimento do Nota Conforme, um sistema integrado baseado em Inteligência Artificial (IA) que utiliza técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (ML) para classificação automática de serviços descritos na discriminação da Nota Fiscal de Serviços Eletrônica (NFS-e). A proposta surge da necessidade de otimizar auditorias fiscais e combater a evasão tributária decorrente da declaração de serviços em desacordo com as alíquotas aplicáveis. A metodologia envolveu a coleta e rotulação de NFS-e, pré-processamento textual, representação vetorial com TF-IDF e a avaliação comparativa de sete algoritmos de classificação, resultando na escolha do *Random Forest* como modelo final. Para viabilizar sua aplicação prática, o modelo foi encapsulado em uma API, possibilitando a integração com sistemas fiscais existentes. Os resultados demonstraram alta acurácia na classificação, evidenciando o potencial da solução para apoiar auditores na identificação de inconsistências e no aprimoramento dos processos de fiscalização tributária. Conclui-se que a aplicação de IA na classificação de serviços descritos em NFS-e representa uma abordagem promissora para ampliar a eficiência da arrecadação e mitigar perdas tributárias.

Palavras-chave: Inteligência Artificial, Processamento de Linguagem Natural, Aprendizado de Máquina, Nota Fiscal de Serviços Eletrônica, Auditoria Fiscal.

ABSTRACT

This study presents the development of Nota Conforme, an integrated system based on Artificial Intelligence (AI) that employs Natural Language Processing (NLP) and Machine Learning (ML) techniques for the automatic classification of services described in the Electronic Service Invoice (NFS-e). The proposal arises from the need to optimize tax audits and address tax evasion resulting from the misreporting of services under incorrect tax rates. The methodology involved the collection and labeling of NFS-e data, text preprocessing, vector representation using TF-IDF, and the comparative evaluation of seven classification algorithms, which led to the selection of Random Forest as the final model. To enable practical application, the model was encapsulated in an API, allowing integration with existing tax systems. The results demonstrated high classification accuracy, highlighting the potential of the solution to assist auditors in identifying inconsistencies and enhancing tax inspection processes. The study concludes that the application of AI to the classification of services described in NFS-e represents a promising approach to improving revenue collection efficiency and mitigating tax losses.

Keywords: *Artificial Intelligence, Natural Language Processing, Machine Learning, Electronic Service Invoice, Tax Audit.*

LISTA DE ABREVIATURAS E SIGLAS

ABRASF	Associação Brasileira das Secretarias de Finanças das Capitais
API	*Application Programming Interface*
CNAE	Classificação Nacional de Atividades Econômicas
CNN	Redes Neurais Convolucionais (do inglês*, Convolutional neural network*)
CSV	Valores Separados por Vírgula (do inglês, *Comma-separated Values*)
CRUD	*Create, Read, Update e Delete*
DT	*Decision Tree*
DW	*Data Warehouse*
FN	Falso Negativo
FP	Falso Positivo
GB	*Gigabyte*
GBT	*Gradient Boosting*
HL	*Hamming Loss*
HTTP	*Hypertext Transfer Protocol*
IA	Inteligência Artificial
IBPT	Instituto Brasileiro de Planejamento e Tributação
IDF	*Inverse Document Frequency*
INPI	Instituto Nacional da Propriedade Industrial
ISS	Imposto Sobre Serviços
JWT	*JSON Web Token*
KB	*Kilobyte*
LGBM	*LightGBM*
LLM	*Large Language Models*
LOA	Lei Orçamentária Anual
LR	*Logistic Regression*
LSTM	*Long Short-Term Memory*
LSV	*Linear Support Vector*
ML	Aprendizado de Máquina (do inglês, *Machine Learning*)
NBM	*Naive Bayes Multinomial*
NCM	Nomenclatura Comum do Mercosul
NF-e	Nota Fiscal Eletrônica
NFKD	*Normalization Form KC, Compatibility Composition*
NFS-e	Nota Fiscal de Serviço Eletrônica
NLG	Geração de Linguagem Natural (do inglês, *Natural Language Generation*)
NLTK	*Natural Language Toolkit*
NLU	Interpretação de Linguagem Natural (do inglês, *Natural Language Understanding*)
OE	Objetivos Específicos
OPME	Órteses, Próteses e Materiais Especiais
PB	Paraíba
PE	Pernambuco
PLN	Processamento de Linguagem Natural
QP	Questão de Pesquisa
RNN	Redes Neurais Recorrente (do inglês, *Recurrent Neural Network*)
RF	*Random Forest*

RS	Rio Grande do Sul
SBBD	Simpósio Brasileiro de Banco de Dados
SEFAZ	Secretaria da Fazenda
SPED	Sistema Público de Escrituração Digital
SQL	Linguagem de Consulta Estruturada (do inglês, *Structured Query Language*)
SUS	Sistema Único de Saúde
SVM	*Support Vector Machines*
TAM	*Technology Acceptance Model*
TMP	Tempo Médio de Predição
TMT	Tempo Médio de Treinamento
TF	*Term Frequency*
TF-IDF	*Term Frequency-Inverse Document Frequency*
UEPB	Universidade Estadual da Paraíba
UPA	Unidade de Pronto Atendimento
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Este capítulo visa contextualizar o objeto pesquisado, apresentando as motivações, problemas de pesquisa, justificativa, objetivos gerais e específicos, metodologia, além da organização da dissertação.

1.1. Motivação

Para estreitar a relação entre o fisco e seus contribuintes, o Brasil inspirou-se em experiências de países como Espanha, Chile e México para instituir o Sistema Público de Escrituração Digital (SPED), como parte do processo de modernização da gestão pública (GERON et al., 2011, p. 47). Instituído pelo Decreto n.º 6.022, de 22 de janeiro de 2007, o SPED é um instrumento para integração de documentos relacionados à escrituração contábil e fiscal de empresários e pessoas jurídicas a partir de fluxo único e computadorizado, centralizando as atividades de recepção, validação, armazenamento e autenticação dos documentos fiscais (BRASIL, 2007).

Segundo Silva et al. (2013, p. 447), o SPED é uma iniciativa conjunta dos governos federal, estadual e municipal, composta por três projetos principais: Escrituração Contábil Digital (SPED Contábil), Escrituração Fiscal Digital (SPED Fiscal) e Nota Fiscal Eletrônica (NF-e). A NF-e foi o primeiro projeto a ser implementado dentro desse contexto (GERON et al., 2011, p. 50). Contudo, a NF-e visa documentar as operações de compras e vendas de mercadorias, cujo controle tributário é de responsabilidade dos Estados e o armazenamento é centralizado na Receita Federal (SEBRAE, 2023).

Já a Nota Fiscal de Serviços Eletrônica (NFS-e), documento fiscal que norteia esta pesquisa, tem a finalidade de registrar as operações de prestação de serviços (ABRASF, 2008, p. 4). O projeto da NFS-e foi concebido através da iniciativa da Associação Brasileira dos Secretários de Finanças das Capitais (ABRASF), com intuito de desenvolver um modelo conceitual para padronizar o desenvolvimento de sistemas de NFS-e, visando atender a cooperação celebrada entre a União, Estados, Distrito Federal e Municípios para a implantação da NFS-e em território nacional (ABRASF, 2008, p. 3-4). Apesar da existência do modelo nacional desenvolvido pela ABRASF, os municípios podem exercer sua autonomia para implementar seus próprios sistemas (DE ANGELI NETO; MARTINEZ, 2016, p. 52).

O município do Recife–PE, por exemplo, é signatário do modelo desenvolvido pela ABRASF (RECIFE, 2017, p. 3). A implantação da NFS-e no município, instituída pela Lei Municipal n.º 17.407/2008 (RECIFE, 2008c) e regulamentada pelo Decreto Municipal n.º 23.675 de 30 de maio de 2008 (RECIFE, 2008a), centralizou a emissão e o armazenamento eletrônico do documento fiscal em sistema próprio da prefeitura. A adoção dessa nova metodologia, de acordo com De Angeli Neto e Martinez (2016, p. 61), torna os processos de arrecadação e cobrança de impostos relativos à prestação de serviços mais ágeis, aprimorando sobremaneira a administração tributária.

No entanto, para burlar o regramento fiscal, empresas utilizam inúmeros mecanismos para recolher menos impostos (DIAS; BECKER, 2017, p. 1), comprometendo a arrecadação fiscal e, por consequência, causando prejuízos ao erário. De acordo com informações relatadas pelos auditores em reunião com a participação do autor dessa dissertação, um erro cometido com frequência pelos contribuintes, é informar um código de serviço prestado com alíquota de tributação específica e descrever no campo de discriminação da NFS-e serviços distintos do selecionado no campo anterior, que possuem alíquotas diferentes.

Isso ocorre, pois o primeiro campo é uma lista predefina de serviços utilizados para cálculo da tributação. Já o segundo, é um campo de texto livre que tem a finalidade de descrever os serviços prestados e inclusão de outras informações importantes para o prestador e o tomador do serviço (RECIFE, 2024b, p. 42). Segundo Dias e Becker (2017, p. 2), os dois atributos são essenciais para avaliar se arrecadação da nota fiscal foi realizada corretamente.

Nesse contexto, além de gerar dispêndios à Fazenda Pública, a declaração falsa ou omissão de qualquer informação com o intuito de eximir-se, total ou parcialmente, do pagamento de tributos, taxas ou quaisquer adicionais previstos em lei constitui crime de sonegação fiscal no Brasil (BRASIL, 1965).

1.2. Problema de pesquisa

A sonegação fiscal, além de violar princípios constitucionais, prejudica o desenvolvimento do país, eleva a carga tributária e compromete investimentos públicos em áreas essenciais como saúde, educação, segurança e infraestrutura (MACEDO; DINIZ FILHO, 2019, p. 110). Em vista disso, Pinheiro e Cunha (2003, p. 34), informam que, através do monitoramento e análise dos dados de contribuintes, a auditoria fiscal tem como finalidade observar o cumprimento das obrigações tributárias. Dessa forma, a auditoria configura-se como

um instrumento essencial no enfrentamento de fraudes tributárias e na redução dos danos causados ao erário.

Segundo Lins Neto (2021, p. 16-17), por possuir detalhes importantes a respeito dos serviços prestados, a análise do campo de discriminação da NFS-e é essencial em auditorias fiscais por permitir verificar se o serviço descrito corresponde ao tipo de tributação aplicada. Quando realizada manualmente, essa análise requer o cruzamento entre a descrição textual do serviço e a alíquota declarada, possibilitando a identificação de inconsistências em relação ao regramento fiscal.

No entanto, desde sua implantação até julho de 2024, o município do Recife–PE registrou aproximadamente 377 milhões de NFS-e emitidas (RECIFE, 2024a). Conforme os dados utilizados nesta pesquisa, somente no primeiro semestre de 2024 foi registrada uma média diária de 82 mil emissões. Assim, o crescente volume de notas emitidas revela a inviabilidade de uma auditoria humana que controle a arrecadação e cobrança dos tributos por meio de métodos tradicionais (DE ANGELI NETO; MARTINEZ, 2016, p. 61).

Por se tratar de um atributo preenchido em linguagem natural livre, não seria possível identificar os serviços descritos na discriminação da NFS-e de forma automatizada apenas com a utilização de técnicas e ferramentas tecnológicas tradicionais, como a execução de instruções em *Structured Query Language* (SQL), dificultando o cruzamento de dados e a identificação de inconsistências em notas fiscais (LINS NETO, 2021, p. 17).

Nesse contexto, o problema central deste estudo reside na limitação de escala enfrentada pelos Auditores do Tesouro Municipal da Prefeitura do Recife–PE, uma vez que, embora tenham capacidade de identificar manualmente os serviços descritos nas NFS-e com elevado grau de precisão, o grande volume de documentos inviabiliza a ampla cobertura dessa verificação, dificultando a análise da conformidade das alíquotas aplicadas.

Diante desse desafio, de acordo com Soares e Cunha (2020, p. 224), a inserção de Aprendizado de Máquina (ML) nas organizações governamentais tem apresentado resultados práticos significativos, ao fornecer ferramentas capazes de analisar, aprender e fazer previsões por meio de dados históricos das bases de dados do governo. Portanto, a partir dos problemas apresentados, este estudo propõe responder às seguintes questões de pesquisa (**QP**).

- Questão 01 (**QP01**): Qual algoritmo de ML oferece o melhor equilíbrio entre precisão e consumo de recurso computacional na tarefa de identificar serviços em NFS-e?
- Questão 02 (**QP02**): A automatização da identificação de serviços em NFS-e, por meio de técnicas de ML, contribui para ampliar a escalabilidade e a cobertura das atividades de fiscalização, integrando-se com facilidade às rotinas e processos de fiscalização realizados pelos auditores?

A primeira questão visa avaliar, com base em métricas de específicas, a eficácia de diferentes modelos de aprendizado de máquina na classificação dos serviços descritos nas NFS-e, de modo a subsidiar a escolha do modelo mais adequado para realizar previsões confiáveis. A segunda investigação busca avaliar se a automatização da identificação de serviços em NFS-e, por meio de técnicas de ML, tem potencial para aumentar a produtividade das fiscalizações, ampliando a escalabilidade e a cobertura das notas analisadas pelos auditores. Além disso, procura-se assegurar que a solução possa ser facilmente incorporada às rotinas e processos já adotados por auditores fiscais em prefeituras e órgãos de controle.

Com as respostas obtidas neste estudo, espera-se desenvolver um sistema que auxilie os auditores a tornarem as fiscalizações mais eficientes, contribuindo para a mitigação dos prejuízos aos cofres públicos e para a inibição das práticas de sonegação, as quais representam um verdadeiro desastre social, ético, público e econômico para a sociedade (MACEDO; DINIZ FILHO, 2019, p. 109).

1.3. Justificativa

A evasão fiscal é um fenômeno global que impacta negativamente toda a sociedade (SOARES; CUNHA, 2020, p. 223). No Brasil, estudo do Instituto Brasileiro de Planejamento e Tributação (IBPT, 2023) estimou em R$ 374 bilhões o valor sonegado anualmente, com indícios de irregularidade em 47% das pequenas empresas, 31% das médias e 16% das grandes. Para efeito de comparação, esse valor é suficiente para cobrir todo orçamento do Ministério da Saúde em 2024, fixado em R$ 232,05 bilhões pela Lei Orçamentária Anual (LOA) (BRASIL, 2024). Além disso, permitiria a construção de três Unidades de Pronto Atendimento (UPA) de Porte III em cada um dos 5.570 municípios do país, considerando o custo médio de R$ 7,8 milhões por unidade (FNS, 2023).

Nesse cenário, os governos buscam exercer um controle crescente sobre seus sistemas de administração tributária para combater a prática de sonegação fiscal (GERON et al., 2011, p. 47). Em razão disso, a identificação de contribuintes com maior potencial de risco fiscal é uma das principais tarefas das administrações públicas (SOARES; CUNHA, 2020, p. 224). Para Dias e Becker (2017, p. 2), devido à inviabilidade de uma fiscalização manual e individualizada por conta do excesso de empresas e o expressivo volume de notas fiscais emitidas diariamente, os auditores utilizam suas experiências e conhecimentos para definição de critérios com objetivo de selecionar as amostras a serem auditadas. Esse processo, embora útil, apresenta cobertura limitada e elevado risco de subnotificação de fraudes.

Segundo Dos Anjos e Pinheiro (2024, p. 185), as auditorias tributárias podem ser beneficiadas com a implementação de soluções de IA em seus processos. Nesse contexto, de acordo com Huyen (2023, p. 4-6), técnicas de ML podem ser particularmente adequadas para problemas que envolvem grandes volumes de dados e padrões complexos, como ocorre em tarefas envolvendo análise de NFS-e.

Dessa forma, este estudo justifica-se pelo potencial de aplicar técnicas de ML na automatização da classificação dos serviços descritos em NFS-e, possibilitando ampliar o volume de documentos analisados, apoiar os auditores na identificação de possíveis fraudes fiscais e contribuir para uma fiscalização mais eficiente da arrecadação tributária.

1.4. Objetivos

1.4.1. Objetivo Geral

Propor e avaliar uma solução de Inteligência Artificial (IA) para a classificação automática dos serviços descritos na Nota Fiscal de Serviços Eletrônica (NFS-e), contemplando etapas de pré-processamento de texto, aplicação de técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (ML), de modo a selecionar o modelo de melhor desempenho e disponibilizá-lo por meio de uma *Application Programming Interface* (API). Busca-se, com isso, ampliar a cobertura, a escalabilidade e a rapidez na análise das notas fiscais, oferecendo suporte aos auditores na detecção de possíveis inconsistências.

1.4.2. Objetivos específicos

Para alcançar o objetivo principal deste estudo, foram estabelecidos os seguintes objetivos específicos (**OE**):

- **OE1:** Coletar, rotular e tratar o conjunto de dados contendo NFS-e do setor de saúde disponíveis no Data Warehouse (DW) de NFS-e;
- **OE2:** Investigar**,** avaliar e selecionar algoritmos de ML para classificação de serviços descritos na discriminação da NFS-e;
- **OE3:** Desenvolver e validar um modelo de ML multirrótulo para classificação de serviços descritos na discriminação da NFS-e;
- **OE4:** Desenvolver uma API para encapsular o modelo preditivo, visando à automatização das predições sobre descrições de serviços na NFS-e e à disponibilização do serviço de inferência em ambiente de produção;
- **OE5:** Avaliar o desempenho do sistema na execução do seu fluxo de trabalho;
- **OE6:** Avaliar a aceitação do sistema por meio da análise das percepções relatadas pelos potenciais usuários.

1.5. Metodologia

Este estudo visa solucionar um problema prático por meio da aplicação de técnicas de IA, com foco em PLN e ML, utilizando dados provenientes de NFS-e. Por essa razão, a pesquisa é caracterizada como aplicada. Esse tipo de investigação é para produzir conhecimentos com foco na aplicação prática, voltados para a resolução de problemas concretos e relacionados a contextos e interesses locais (PRODANOV; FREITAS, 2013, p. 51).

A presente pesquisa adota uma abordagem mista, combinando métodos quantitativos e qualitativos. A abordagem quantitativa, segundo Prodanov e Freitas (2013, p. 69-70), é caracterizada pelo uso de dados numéricos e pela aplicação de procedimentos estatísticos para testar hipóteses e verificar relações entre variáveis. No contexto deste estudo, ela está presente na etapa experimental, que envolve a aplicação e comparação de diferentes algoritmos de IA com base em critérios objetivos, como assertividade e desempenho computacional, a fim de identificar a solução mais eficiente para o problema proposto.

Já a abordagem qualitativa, entende a realidade como inseparável da subjetividade do sujeito, priorizando a interpretação de significados em contextos naturais, sem uso de métodos estatísticos (PRODANOV; FREITAS, 2013, p. 70). Neste trabalho, a dimensão qualitativa refere-se a análise das percepções dos potenciais usuários sobre a viabilidade, facilidade e intenção de uso do sistema desenvolvido, além de seu potencial para melhorar os processos de fiscalização. Quanto ao método, trata-se de uma pesquisa de natureza experimental, uma vez

que envolve a aplicação prática de técnicas de IA em um ambiente controlado, visando testar hipóteses e avaliar os resultados obtidos a partir da manipulação de variáveis específicas (WAZLAWICK, 2021).

A metodologia adotada compreende cinco etapas principais: (i) coleta, preparação e rotulação de dados reais de NFS-e extraídos do DW; (ii) experimentação com diferentes algoritmos de ML, utilizando métricas como acurácia, precisão, *recall*, *F1-score* e *Hamming Loss* para avaliação; (iii) desenvolvimento de um modelo de classificação multirrótulo baseado em técnicas de PLN, como *Term Frequency-Inverse Document Frequency* (TF-IDF); (iv) implementação de uma API para disponibilização do serviço de inferência em ambiente de produção; e (v) apresentação do sistema Nota Conforme e coleta da percepção dos potenciais usuários.

1.6. Organização da dissertação

Com propósito de abordar todos os aspectos do estudo de forma estruturada e lógica, este documento foi organizado em sete capítulos. O **Capítulo 1** apresenta a introdução, que contextualiza a pesquisa, apresentado os fundamentos iniciais do estudo realizado, incluindo: motivação, problemática, justificativa, objetivos gerais e específicos, metodologia, além da organização desta dissertação. Já no **Capítulo 2** é realizada a fundamentação teórica, que visa explorar os principais conceitos que dão embasamento teórico para o estudo, incluindo os fundamentos do PLN e de ML, método TF-IDF para representação de textos e métricas de avaliação de modelos. No **Capítulo 3** são apresentados os trabalhos relacionados, discute pesquisas previamente desenvolvidas com propostas similares, e possuem relevância para este estudo. No **Capítulo 4** são apresentados os materiais e métodos utilizados no processo de desenvolvimento do Nota Conforme. Descreve detalhadamente o processo experimental e as técnicas empregadas no estudo, tais como: estratégias para coleta e rotulação da amostra; técnicas de limpeza e preparação dos dados com o pré-processamento; extração de características do texto por meio da engenharia de recursos; avaliação comparativa dos modelos; implementação dos modelos classificadores; desenvolvimento da API como proposta para implantação do modelo em ambiente de produção. No **Capítulo 5** o sistema Nota Conforme é apresentado através dos diagramas arquiteturais, além da exploração de suas funcionalidades. Já o **Capítulo 6** apresenta os resultados e discussões, que discute os achados da pesquisa por meio das avaliações que nortearam a escolha do modelo final, a avaliação de desempenho da API, evidenciando sua viabilidade de implantação e as percepções dos

potenciais usuários quanto a sua utilidade, facilidade e potencial de uso no futuro. Por fim, no **Capítulo 7 s**ão discutidas as considerações finais, que resume os principais achados da pesquisa, destacando as contribuições e limitações do estudo, além de propor direções para trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Esta seção revisitará a literatura para apresentar as técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (ML) aplicadas no desenvolvimento do classificador proposto.

2.1. Processamento de Linguagem Natural

O PLN corresponde a um campo da Inteligência Artificial (IA) que se propõe a investigar e desenvolver técnicas de processamento computacional da linguagem humana, tanto escrita quanto falada (CASELI; NUNES, 2024, p. 10). De acordo com Russell e Norvig (2022, p. 745), os três principais objetivos do PLN por meio de computadores são: possibilitar a comunicação com os seres humanos, permitir o aprendizado a partir do vasto volume de textos produzidos pelo ser humano e promover avanços dos conhecimentos científicos das linguagens e de seu uso.

Segundo Caseli e Nunes (2024, p. 10), o PLN é dividido em duas grandes subáreas. A primeira é responsável pela tarefa de análise e interpretação da língua humana, denominada de Interpretação de Linguagem Natural, ou em inglês *Natural Language Understanding* (NLU). A segunda, definida como Geração de Linguagem Natural, ou em inglês, *Natural Language Generation* (NLG), por sua vez, visa a geração de linguagem humana. De acordo com Vajjala et al. (2020, p. 6-7), novas aplicações do PLN surgem cotidianamente, mas destacam que as tarefas fundamentais do PLN incluem:

- **Modelagem de linguagem:** Prever a próxima palavra em uma frase, considerando o contexto formado pelas palavras anteriores;
- **Classificação de texto:** Agrupar textos em categorias com base em seu conteúdo. Essa é considerada uma das tarefas mais populares em PLN;
- **Extração de informações:** Realizar a extração de informações relevantes de um texto, como, por exemplo, identificar pessoas mencionadas em uma postagem na rede social;
- **Recuperação de informações:** Localizar documentos relevantes com base nas consultas dos usuários. A pesquisa no Google é um exemplo clássico desse tipo de tarefa;
- **Agente de conversação:** Interagir com o ser humano na sua língua nativa, como fazem os assistentes virtuais Alexa, Siri e Google Assistente;

- **Resumo de texto:** Gerar resumos de textos mais longos, mantendo seu sentido e significado principais;
- **Resposta à pergunta:** Responder automaticamente às perguntas formuladas em linguagem natural pelos usuários;
- **Tradução automática:** Traduzir texto de um idioma para outro;
- **Modelagem de tópicos:** Identificar e extrair tópicos de grandes coleções de documentos, permitindo reconhecer os temas predominantes nos textos.

Nesta pesquisa, dada a natureza predominantemente textual, o uso de PLN será fundamental no processamento da discriminação da Nota Fiscal de Serviço Eletrônica (NFS-e) para identificação dos serviços descritos. Como se trata de um problema de classificação, o PLN será combinado com técnicas de ML, que serão detalhadas na próxima seção.

2.2. Aprendizado de Máquina

O ML "é a ciência (e arte) da programação de computadores de modo que eles possam aprender com os dados" (GÉRON, 2021, p. 3). Segundo Russel e Norvig (2022, p. 590), no aprendizado de máquina o modelo é desenvolvido a partir da observação dos dados de treinamento e, em seguida, é utilizado para realizar novas previsões. Para Kalinowski et al. (2023, p. 273), o aprendizado pode ser dividido em quatro categorias: aprendizado supervisionado, não supervisionado, semi-supervisionado e por reforço.

- **Aprendizado supervisionado**: O algoritmo recebe um conjunto de dados de exemplos com pares de entrada e saída para treinamento e aprende uma função que realiza o mapeamento entre elas (RUSSEL E NORVIG, 2022, p. 591). Isso significa que o conjunto de dados previamente rotulado deve ser submetido ao algoritmo, que gerará um modelo capaz de fazer predições do valor do atributo alvo em novos objetos a partir da observação de seus atributos preditivos (FACELI et al., 2023, p. 3).
- **Aprendizado não supervisionado**: De acordo com Kalinowski et al. (2023, p. 273), o processo de aprendizado não supervisionado busca identificar regularidades nos dados para agrupá-los com base nas similaridades que apresentam entre si, pois não há informação sobre a saída desejada.
- **Aprendizado semi-supervisionado**: O aprendizado semi-supervisionado ocorre quando os dados estão parcialmente rotulados. Normalmente, a maioria das instâncias estão sem rótulos, possuindo poucas instâncias rotuladas (GÉRON, 2021, p. 12). Ou

seja, é uma técnica proveniente da combinação da supervisionada e não supervisionada (KALINOWSKI et al., 2023, p. 273).

- **Aprendizado por reforço**: Segundo Russel e Norvig (2022, p. 592), no aprendizado por reforço, o agente obtém conhecimento por meio de uma sequência de reforços, que incluem recompensas e punições. É função do agente definir quais ações executadas anteriormente foram determinantes para os resultados obtidos e modificar suas ações visando conquistar mais recompensas futuramente.

Esta pesquisa, no entanto, utilizará somente o aprendizado supervisionado. Esse tipo de aprendizado pode ser dividido em dois tipos de tarefas de predição: regressão e classificação (FACELI et al., 2023, p. 3). De acordo com Russel e Norvig (2022, p. 591), em problemas de regressão a saída do modelo será sempre um número, como a previsão da temperatura do dia seguinte. Para Kalinowski et al. (2023, p. 276), quando a saída for categórica, também conhecida como classe, trata-se de um problema de classificação. Rotular verdadeiro ou falso para identificar se um e-mail é *spam*, por exemplo, é uma tarefa de classificação.

Segundo Vajjala et al. (2020, p. 120), o problema de classificação pode ser subdivido em três tipos, com base no número de classes possíveis: classificação binária, multiclasse e multirrótulo. Nas classificações binária e multiclasse, cada documento pode ser associado a somente uma classe dentre todas as possíveis. A classificação binária é caracterizada pela existência de somente duas classes. Conforme exemplo ilustrado na Tabela 1, a partir dos sintomas fornecidos ao modelo, um paciente é classificado como doente ou saudável.

Tabela 1 - Exemplo de representação de uma classificação binária.

PACIENTE	TOSSE	FEBRE	FALTA DE AR	MANCHAS VERMELHAS	RESULTADO (CLASSE)
1	Sim	Sim	Não	Não	Doente
2	Não	Não	Não	Não	Saudável
3	Sim	Sim	Sim	Não	Doente
4	Não	Sim	Sim	Sim	Doente
5	Não	Não	Não	Não	Saudável

Fonte: Elaborado pelo autor

A classificação multiclasse, por outro lado, ocorre ao ter mais de duas classes de saída, como, por exemplo, uma análise de sentimento que classifica o cliente como positivo, neutro ou negativo (VAJJALA et al., 2020, p. 120). Conforme o exemplo apresentado na Tabela 2, o modelo prever a doença do paciente a partir dos sintomas informados, podendo o paciente estar

com Gripe, Covid-19, Dengue ou até mesmo saudável. Embora o modelo possua quatro classes de saída possíveis, cada paciente só pode ser associado a somente uma delas.

Tabela 2 - Exemplo de representação de uma classificação multiclasse.

PACIENTE	TOSSE	FEBRE	FALTA DE AR	MANCHAS VERMELHAS	RESULTADO (CLASSE)
1	Sim	Sim	Não	Não	Gripe
2	Não	Não	Não	Não	Saudável
3	Sim	Sim	Sim	Não	Covid-19
4	Não	Sim	Sim	Sim	Dengue
5	Não	Não	Não	Não	Saudável

Fonte: Elaborado pelo autor

Já no problema de classificação multirrótulo, um documento pode ser associado simultaneamente a mais de um rótulo. A Tabela 3 efetua a representação de um resultado hipotético de uma classificação multirrótulo. É possível observar, por exemplo, que os objetos de ID 1 e 3 possuem, respectivamente, dois e três serviços em suas descrições, sendo atribuídos ao último todos os rótulos possíveis.

Tabela 3 - Exemplo de representação de uma classificação multirrótulo.

ID	DISCRIMINAÇÃO DA NFS-e	CONSULTA MÉDICA	EXAME LABORATORIAL	EXAME DE IMAGEM
1	Consulta com ortopedista e radiografia do joelho	1	0	1
2	Exames laboratoriais: hemograma, colesterol, glicemia	0	1	0
3	Consulta médica, raio-x do tórax e hemograma completo	1	1	1
4	Consulta médica com clínico geral	1	0	0

Fonte: Elaborado pelo autor

Esses conceitos podem ser aplicados a uma variedade de tipos de dados, como imagens, planilhas, documentos e banco de dados (KALINOWSKI et al., 2023, p. 20). A classificação multirrótulo é comumente utilizada em problemas de classificação de textos. Isso ocorre porque, em muitos casos, um texto pode abordar múltiplos temas ou categorias, o que torna essa abordagem especialmente útil (FACELI et al., 2023, p. 264).

2.2.1. Representação de textos

É factível ao ser humano interpretar manualmente o conteúdo de um texto e atribuir a ele uma categoria que o represente melhor a partir de sua leitura. Entretanto, a realização dessa tarefa manualmente se torna inviável ao se trabalhar com grandes volumes de dados, necessitando a utilização de estratégias para automatização (FACELI et al., 2023, p. 308). Para isso, no contexto do problema de classificação existe categorização de textos. Nessa abordagem, o atributo de entrada (preditivo) é do tipo texto e o objetivo é categorizá-lo em uma ou mais classes. O comprimento do atributo de entrada é indefinido, podendo ser somente uma letra, palavra, frase, parágrafo ou até mesmo um documento inteiro (VAJJALA et al., 2020, p. 119).

No entanto, os computadores não possuem capacidade para interpretá-los em seu formato original. As técnicas de aprendizado de máquina lidam geralmente com informações estruturadas (FACELI et al., 2023, p. 310). Por esse motivo, para ser possível a utilização de dados textuais em tarefas de aprendizado de máquina é necessário convertê-los para alguma representação matemática, esse processo é conhecido como representação de textos (VAJJALA et al. 2020, p. 81-84).

Vajjala et al. (2020, p. 85-92) abordam algumas técnicas que utilizam vetores números para representar os textos, tais como: *One-Hot Encoding*, *Bag of Words*, *Bag of N-Grams*. Essas abordagens não consideram que determinadas palavras são mais importantes do que outras, considerando que todas as palavras são igualmente relevantes. Como alternativa a essa limitação, os autores consideram a utilização da técnica *Term Frequency-Inverse Document Frequency* (TF-IDF), que mensura a importância de determinada palavra em relação as demais no documento e no *corpus*.

2.2.2. Term Frequency-Inverse Document Frequency (TF-IDF)

A abordagem TF-IDF é um algoritmo que utiliza métodos estatísticos para quantificar a relevância de uma palavra para um documento ou coleção de documentos. Uma palavra é considerada relevante para identificar uma categoria quando ela aparece com frequência em determinado documento, mas raramente aparecem no conjunto de documentos. A técnica é constituída de duas partes: frequência do termo e frequência inversa de documentos (LIU, 2018, p. 218).

A Frequência do Termo, em inglês *Term Frequency* (TF), é uma medida que avalia a frequência que um termo aparece em um documento específico. Em virtude da possibilidade do documentos possuírem diferentes tamanhos em um *corpus*, podendo um termo aparecer com mais frequência em documentos mais longos do que em documentos mais curtos, a divisão do número de ocorrências é realizada pelo comprimento do documento, a fim de normalizar a contagem. Nesse sentido, o resultado de TF de um termo em um documento é definido segundo a equação a seguir (VAJJALA et al. 2020, p. 90):

$$TF(t,d) = \frac{(\text{Número de ocorrências do termo } t \text{ no documento } d)}{(\text{Número total de termos no documento } d)}$$

A Frequência Inversa de Documentos, em inglês *Inverse Document Frequency* (IDF), avalia a importância do termo em um *corpus*. Ele considera que os termos que são muito frequentes em um *corpus* não são tão importantes para identificar determinado documento. Consequentemente, diminui os pesos daqueles que aparecem em muitos documentos e aumenta daqueles que aparecem em poucos documentos. O cálculo de IDF é representado pela equação a seguir (VAJJALA et al. 2020, p. 91):

$$IDF(t) = \log_e\left(\frac{(\text{Número total de documentos no } corpus)}{(\text{Número de documentos que contêm o termo } t)}\right)$$

Logo, o TF avalia a frequência da palavra em determinado documento e o IDF busca equilibrar o nível de importância da palavra a partir da redução do peso de palavras que aparecem em muitos documentos, avaliando sua importância para o conjunto de documentos. Desse modo, segundo Vajjala et al. (2020, p. 91), o produto dessas duas partes constitui o valor de TF-IDF, conforme ilustrado na equação a seguir:

$$\text{TF-IDF} = TF * IDF$$

Para fornecer uma análise mais detalhada dos resultados obtidos ao utilizar a técnica TF-IDF, foi criado um conjunto de dados hipotético, apresentado na Tabela 4. O conjunto de dados consiste em 4 documentos, totalizando 15 termos, dos quais 10 são distintos. No exemplo em questão, considera-se que o conjunto de dados já tenha passado pelas etapas de pré-processamento, com todo o texto convertido para minúsculo, letras acentuadas substituídas por suas respectivas formas sem acentuação, com *stop words* removidos e outras tarefas de uso comum para tratamento de dados.

Tabela 4 - Conjunto de dados hipotético para calcular TF-IDF.

ID_DOCUMENTO	DOCUMENTO
D1	referente consulta medica
D2	referente exame laboratorial hemograma
D3	referente consulta ginecologica
D4	referente consulta dermatologista limpeza pele

Fonte: Elaborado pelo autor

A Tabela 5 apresenta os resultados dos cálculos de TF e IDF individualmente, seguido do valor final de TF-IDF de todos os termos existentes no *corpus*, mostrando sua importância em cada um dos documentos do conjunto de dados. Observa-se nos resultados que foi atribuído o valor de importância zero para o termo "referente", pois ele aparece em todos os documentos do conjunto de dados e não possui relevância para identificar um documento específico dentro do *corpus*. Algo semelhante acontece com a palavra "consulta", que aparece em 3 dos 4 documentos. Apesar de não ser atribuído valor zero, sua importância é bastante reduzida em todos os documentos em que está presente.

Tabela 5 - Resultado do experimento de cálculo do TF-IDF.

TERMOS	TF				IDF	TF-IDF			
	D1	D2	D3	D4		D1	D2	D3	D4
consulta	0,3333	0,0000	0,3333	0,2000	0,2877	0,0959	0	0,0959	0,0575
dermatologista	0,0000	0,0000	0,0000	0,2000	1,3863	0	0	0	0,2773
exame	0,0000	0,2500	0,0000	0,0000	1,3863	0	0,3466	0	0
hemograma	0,0000	0,2500	0,0000	0,0000	1,3863	0	0,3466	0	0
laboratorial	0,0000	0,2500	0,0000	0,0000	1,3863	0	0,3466	0	0
limpeza	0,0000	0,0000	0,0000	0,2000	1,3863	0	0	0	0,2773
medica	0,3333	0,0000	0,0000	0,0000	1,3863	0,4621	0	0	0
ginecologica	0,0000	0,0000	0,3333	0,0000	1,3863	0	0	0,4621	0
pele	0,0000	0,0000	0,0000	0,2000	1,3863	0	0	0	0,2773
referente	0,3333	0,2500	0,3333	0,2000	0,0000	0	0	0	0

Fonte: Elaborado pelo autor

As demais palavras que estão presentes em documentos específicos e não aparecem frequentemente em diferentes documentos do conjunto de dados como, por exemplo, exame, hemograma, laboratorial e nutricionista possuem valores de relevância maiores em seus respectivos documentos. No entanto, esse valor é igual a zero nos documentos nos quais esses termos não aparecem. Por conseguinte, essa análise prática corrobora com o que foi discutido

ao longo desta sessão sobre a abordagem TF-IDF, evidenciando sua capacidade em quantificar a importância de uma palavra em determinado contexto.

Apesar de amplamente utilizado, o TF-IDF apresenta limitações importantes. Por tratar os termos de maneira independente e ignorar a ordem das palavras, a técnica não captura as relações semânticas e contextuais. Além disso, ao gerar vetores esparsos e de alta dimensionalidade, o modelo pode resultar em maior custo computacional e ser mais suscetível a ruídos. Essas limitações evidenciam a importância de investigar, em estudos futuros, abordagens baseadas em *word embeddings*, que representam palavras de forma densa e contextualizada (VAJJALA et al. 2020).

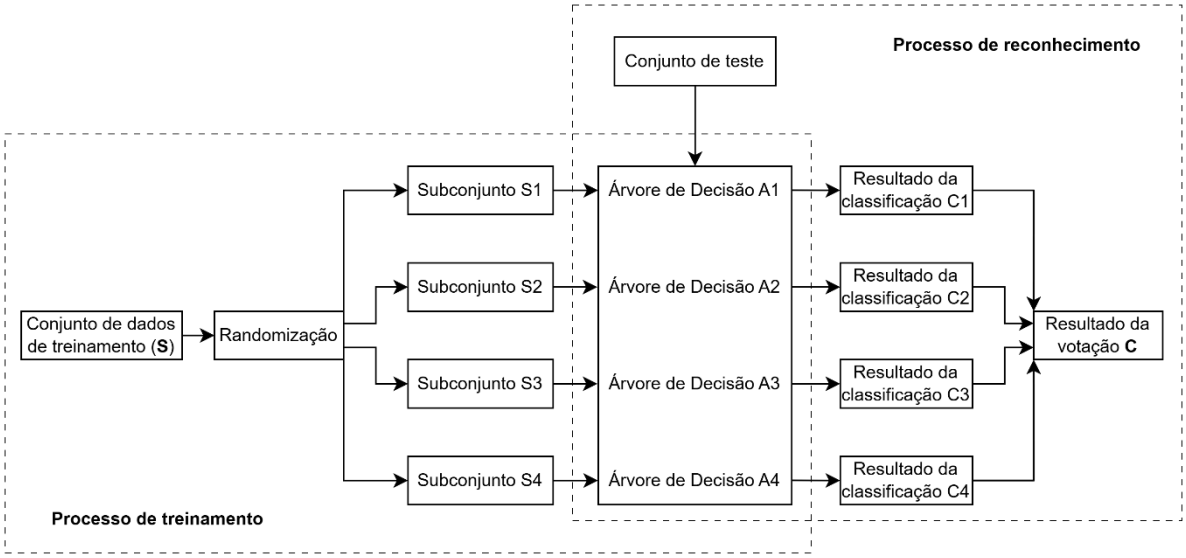
Portanto, existem diversas técnicas e abordagens que viabilizam o desenvolvimento de modelos capazes de categorizar documentos textuais. Nesta pesquisa, foi realizada uma avaliação experimental comparativa entre diferentes algoritmos de aprendizado de máquina utilizando a técnica TF-IDF. A escolha do algoritmo foi realizada considerando os critérios de qualidade preditiva e tempos de treinamento e predição. Com base nos experimentos realizados, o algoritmo *Random Forest Classifier*, que será detalhado na próxima seção, destacou-se como a opção mais eficaz para resolução deste problema.

2.2.3. Random Forest Classifier

O classificador *Random Forest*, ou Floresta Aleatória, foi proposto pela primeira vez por Leo Breiman em seu artigo intitulado de *Random Forest*. As florestas aleatórias são formadas a partir da combinação de diversas árvores de decisão (BREIMAN, 2001, p. 5). Segundo Song, Liu e Yang (2015, p. 723), o processo de randomização é inserido na construção do modelo, que inclui a distribuição do subconjunto de exemplos e de características. Esse processo visa assegurar a independência das árvores de decisão, além de obter um modelo com melhor precisão e capacidade de generalização.

Conforme ilustrado pela Figura 1, as árvores são treinadas de forma independente por meio de subconjuntos de dados aleatórios que recebem do conjunto original, esses subconjuntos são distribuídos de maneira igualitária entre as árvores (BREIMAN, 2001, p. 5). Esse tipo de distribuição é realizado pelo método *bagging* e torna cada subconjunto de treinamento independente (SONG; LIU; YANG, 2015, p. 723).

Figura 1 - Estrutura do Classificador Random Forest.

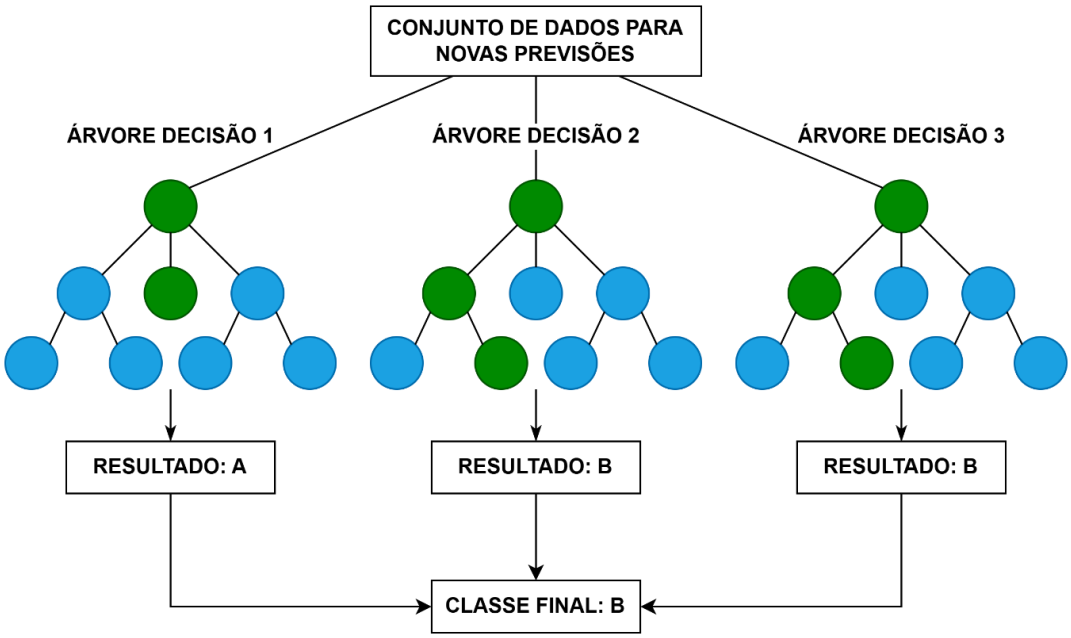


Fonte: (SONG; LIU; YANG, 2015, p. 724)

A seleção de características consiste na escolha aleatória de um subconjunto de todas as características disponíveis, também por meio do método *bagging*. O número de características definido influencia diretamente tanto no desempenho do classificador quanto na correlação entre as árvores. Quanto maior a quantidade de características, melhor o desempenho do modelo. Em contrapartida, há menos diversidade, o que torna as árvores mais semelhantes (SONG; LIU; YANG, 2015, p. 723). Esse número de características do subconjunto pode ser definido por meio de parâmetros específicos como, por exemplo, *max_features* no *Scikit Learn*.

Após a conclusão do treinamento do novo classificador, o resultado que definirá a classe final da previsão de cada um dos novos objetos fornecidos ao modelo será definido com base nos resultados individuais de cada árvore de decisão, conforme ilustrado na Figura 1 (SONG; LIU; YANG, 2015, p. 723). Ou seja, o resultado da previsão será determinado por votação majoritária, isso significa que a classe que receber o maior número de previsões entre as árvores será a classe final, conforme exemplo apresentado na Figura 2.

Figura 2 - Representação de fluxo de novas previsões com Random Forest.



Fonte: Elaborado pelo autor

Segundo Breiman (2004, p. 1), o *Random Forest* mostrou ser um algoritmo preciso e com uma capacidade incomum para lidar com milhares de variáveis sem comprometer seu desempenho ou necessidade de excluí-las. Contudo, apesar de parecer simples, sua análise pode ser complexa e desafiadora. De acordo com Song, Liu e Yang (2015, p. 726), o *Random Forest* é um algoritmo prático e eficaz para resolver problemas de classificação. Parmar, Katariya e Patel (2019, p. 763) reforçam acerca da sua eficiência ao afirmarem que o desempenho do algoritmo sobressai em relação aos classificadores únicos, ao obter maiores taxas na tarefa de classificação. Por esses motivos, o *Random Forest* tem ampliado sua popularidade na comunidade de pesquisa.

2.3. Métricas de avaliação do modelo

Uma etapa essencial de um projeto de ML é a avaliação da qualidade do modelo desenvolvido. A aferição da qualidade do modelo pode ser realizada de diferentes perspectivas, sendo a mais comum a avaliação de assertividade em dados não visto (VAJJALA et al. 2020, p. 68). Segundo Géron (2021, p. 71), esse tipo de avaliação de desempenho pode ser realizado por meio de diversas métricas disponíveis, as quais serão detalhadas ao longo dessa sessão.

A matriz de confusão encontra-se no centro das métricas de avaliação de problemas de classificação (BRUCE; BRUCE, 2019, p. 200). Em razão disso, a sua utilização é fundamental no processo de avaliação de desempenho de um modelo de classificação, ao possibilitar a identificação do número de vezes que as instâncias da classe A são classificadas como classe B (GÉRON, 2021, p. 73). Assim, a matriz de confusão possibilita avaliar o comportamento do modelo comparando a sua saída com os rótulos de referências (JURAFSKY; MARTIN, 2025, p. 66).

A matriz de confusão é formada pelos Verdadeiros Positivos (VP), que representa a quantidade de vezes em que o modelo identificou corretamente uma classe positiva. Por Verdadeiros Negativos (VN), que determina a quantidade de vezes em que o modelo identificou corretamente uma classe negativa. Pelos Falsos Positivos (FP), que indicam quando modelo atribui indevidamente exemplos da classe negativa a classe positiva. Além dos Falsos Negativos (FN), que aponta quando os exemplos da classe positiva são atribuídos indevidamente à classe negativa (FACELI et al., 2023, p. 152). Ou seja, as posições VP e VN representam a quantidade de acertos, enquanto FN e FP corresponde aos erros, conforme ilustrado na Tabela 6.

Tabela 6 - Representação de uma matriz de confusão.

		Classe predita	
		+	–
Classe verdadeira	+	VP	FN
	–	FP	VN

Legenda	
	Acertos
	Erros

Fonte: Adaptado de Faceli et al. (2023, p. 153)

A partir dos resultados obtidos na matriz de confusão, é possível calcular novas métricas que contribuem para a avaliação do modelo de classificação. Entre elas, estão as métricas de Acurácia, Precisão, Revogação (*recall*) e F1-Score, as quais são frequentemente utilizadas para medir o desempenho de modelos em problemas de classificação (BRUCE; BRUCE, 2019, p. 202). A Acurácia, por exemplo, se apresenta como uma das métricas mais utilizadas para avaliar modelos de classificação. Ela encontrará a taxa total de acertos do classificador, podendo ser extraída da matriz de confusão, pois é resultado da divisão do total de exemplos classificados corretamente pela quantidade total de exemplos classificados pelo modelo, representada pela equação (KALINOWSKI et al. 2023, p. 281-282):

$$Acurácia = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN}}$$

Já a Precisão, de acordo com Faceli et al. (2023, p. 153), é a "proporção de exemplos positivos classificados corretamente entre todos aqueles preditos como positivos". Ou seja, avalia quantas vezes o modelo classificou um exemplo como positivo e realmente era. Segundo Kalinowski et al. (2023, p. 283), a Precisão pode ser representada pela equação:

$$Precisão = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

O *Recall*, por sua vez, avalia a capacidade do modelo em identificar todas as instâncias positivas (BRUCE E BRUCE, 2019, p. 202). O *Recall* é resultado da divisão das predições VP pela quantidade total de instâncias positivas no conjunto de dados, conforme apresentada na equação (KALINOWSKI et al. 2023, p. 282-283):

$$Recall = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Adicionalmente, Géron (2021, p. 75) defende a importância de integrar em uma única métrica a precisão e o *recall*. Essa integração resulta numa nova métrica chamada de *F1-score*, calculada por meio da média harmônica da precisão e do *recall*. Ao contrário da média aritmética, que trata todos os valores de maneira igualitária, a média harmônica atribui mais relevância aos valores mais baixos. Dessa forma, o modelo só atingirá um *F1-score* satisfatório se ambas as métricas apresentarem resultados elevados. Assim, a métrica *F1-score* pode ser representada pela equação:

$$\text{F1-score} = 2 \times \frac{Precisão \times Recall}{Precisão + Recall}$$

Além disso, na classificação multirrótulo é possível que um exemplo seja classificado de forma parcialmente correta ou incorreta. Isso ocorre quando o classificador acerta alguns dos rótulos atribuídos a um exemplo, mas omite outros ou, ainda, quando atribui rótulos indevidos. Nesse contexto, o *Hamming Loss* (HL) é uma métrica utilizada para avaliar esses tipos de classificadores, sendo que valores mais baixos indicam melhor desempenho do modelo. A predição perfeita ocorre quando essa métrica atinge o valor zero, o que significa que não houve erros na atribuição dos rótulos (FACELI et al., 2023, p. 269). Segundo os autores, a definição é dada pela equação a seguir:

$$Hamming\ Loss\ \left(\hat{f}, \mathbf{X}\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{a(\mathbf{Y}_i, \mathbf{Z}_i)}{k}$$

A avaliação de um modelo de aprendizado supervisionado pode ser realizada por meio de diversas medidas (KALINOWSKI et al. 2023, p. 281). Em razão disso, é necessário selecionar as métricas de avaliação que se adequem melhor ao problema de aprendizado de máquina e necessidades do projeto desenvolvido. Com vista em obter uma análise detalhada do desempenho do modelo desenvolvimento, todas as métricas vistas nesta sessão serão utilizadas no processo de avaliação do classificador.

3. TRABALHOS RELACIONADOS

O surgimento de pesquisas propondo a utilização de técnicas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina (ML) para identificação de características relevantes em notas fiscais eletrônicas, a fim de mitigar de dispêndios aos cofres públicos a partir da otimização dos processos de auditorias e fiscalizações dos órgãos controladores, contribuíram significativamente para o desenvolvimento desta pesquisa. Desse modo, este capítulo visa explorar essas pesquisas e extrair contribuições com intuito de enriquecer os resultados deste trabalho.

Gomes (2023) propôs a implementação de uma ferramenta baseada em inteligência artificial para identificar e classificar Órteses, Próteses e Materiais Especiais (OPME) a partir da descrição do produto em Notas Fiscais Eletrônicas (NF-e). Para atender os requisitos da pesquisa, foram desenvolvidos três modelos classificadores para cada algoritmo avaliado, executados sequencialmente. O primeiro classifica o tipo de OPME a partir da descrição do produto, aqueles que não são OPME o modelo classifica-os como Outros. O segundo identifica a qual classe pertence o produto, possuindo 83 classes distintas como, por exemplo, Cateter, Prótese Ortopédica, Cânula. Já o terceiro, pretende identificar o possível Procedimento do Sistema Único de Saúde (SUS) utilizado pelo produto.

O pesquisador selecionou uma amostra de 4.549 registros de um total de 465.726 NF-e emitidas entre janeiro de 2020 e maio de 2022, disponibilizados pelo governo do Rio Grande do Norte. A pesquisa avaliou o desempenho de sete algoritmos de classificação. Os resultados revelaram que os algoritmos *Linear Support Vector* e *Random Forest* se destacaram em relação aos demais, obtendo os melhores níveis de desempenho nas avaliações, com resultados semelhantes para os três tipos de classificação. A Tabela 7 apresenta os resultados, evidenciando a acurácia dos algoritmos nas três tarefas de classificação propostas pela pesquisa.

Tabela 7 - Acurácia dos algoritmos Linear Support Vector e Random Forest.

Algoritmos	Tipo	Classe	Procedimento
Linear Support Vector	0,998	0,996	0,986
Random Forest	0,996	0,995	0,974
Decision Tree	0,995	0,992	0,967
Gradient Boosting	0,995	0,989	0,967
Naïve Bayes Bernoulli	0,990	0,735	0,371
Naïve Bayes Multinomial	0,990	0,983	0,893
Naïve Bayes Gaussian	0,990	0,989	0,946

Fonte: Adaptado de Gomes (2023)

Os modelos desenvolvidos foram encapsulados em uma aplicação *Web* para implantação. Por meio da ferramenta, o usuário tem a possibilidade de dar entrada com uma descrição de produto e o sistema efetuará as classificações, exibindo os resultados em seguida. Diferentemente do trabalho de Gomes (2023), que se concentra na classificação de OPME e identificação de procedimentos SUS, o presente estudo propõe um modelo de classificação multirrótulo voltado à identificação dos códigos de serviços prestados descritos em Nota Fiscal de Serviço Eletrônica (NFS-e), considerando as particularidades do controle tributário nos municípios. Além disso, enquanto Gomes (2023) apresenta uma interface web, este trabalho foca na entrega de uma *Application Programming Interface* (API), visando integração direta com os sistemas dos órgãos fiscalizadores.

Já Lins Neto (2021), desenvolveu um sistema baseado em regras explícitas para automatizar parte das atividades desenvolvidas pelos auditores da Prefeitura Municipal de Ipojuca–PE nos processos de auditoria, por meio da extração de informações úteis e classificação de NFS-e suspeitas de fraude do setor da Construção Civil. Através da ferramenta, o usuário pode importar um conjunto de dados de NFS-e em arquivo no formato de Valores Separados por Vírgula (CSV, do inglês *Comma-separated values*) e receber um novo arquivo no mesmo formato com o resultado do processamento das extrações das informações e classificações. Com base nas leis que regulamentam o Imposto Sobre Serviços (ISS) e conhecimentos dos auditores, foram definidas seis regras para extração de informações relevantes da discriminação da NFS-e, conforme Tabela 8.

Tabela 8 - Regras de extração de informações da NFS-e.

ID	Regra	Objetivo
1	Extração do período de prestação do serviço	Extração da data inicial e final da prestação do serviço para cruzamento com data de competência da NFS-e. Essa informação é utilizada nas regras de classificação.
2	Extração do número do contrato	Extração do número de contrato firmado entre a empresa prestadora e a tomadora do serviço. Essa informação não é utilizada para classificação, mas útil para os auditores, por ser obrigatório informá-la para utilização de incentivo fiscal.
3	Extração do número do Relatório de Medição	Extrai somente o número do relatório de medição. Esse relatório informa o planejamento da obra e o que já foi executado. Não é utilizado nas regras de classificação, mas útil para os auditores solicitarem informações adicionais quando necessário.
4	Extração do número da Nota de Liquidação	Extrai o número do recibo, uma nota sem valor fiscal. Não é utilizada nas regras de classificação, mas útil para os auditores solicitarem informações adicionais quando necessário.

5	Extração de valores	Extrai os valores dos serviços quando informados na discriminação da NFS-e. Não é utilizada nas regras de classificação, mas útil para o processo de auditoria.
6	Extração de serviços sem isenção	Extrai nomes dos serviços que não possuem isenção de ISS. Essa informação é utilizada nas regras de classificação.

Fonte: Adaptado de Lins Neto (2021)

Além disso, para identificar NFS-e com possíveis indícios de fraudes para auxiliar os processos de auditoria, foram definidas quatro regras explícitas de classificação, conforme disposto na Tabela 9.

Tabela 9 - Regras de classificação de notas suspeitas de fraudes.

ID	Regra	Objetivo
1	Período da prestação de serviços x Competência da NFS-e	Comparar a data final da prestação de serviço com a data de competência da NFS-e. Em caso de divergência, acumula-se pontuação de fraude.
2	Valor dos serviços x Base de Cálculo	Comparar o valor total de serviços prestados com o valor da base de cálculo. Qualquer divergência acumula pontuação de fraude.
3	Serviços sem isenção de ISS	Verificar se a nota com isenção total de ISS possui algum serviço que não se enquadra na regra de isenção. Acumula pontuação de fraude se positivo.
4	Notas emitidas com isenção fora do período	Comparar a data da emissão da NFS-e com isenção total de ISS com a data final do período do benefício de isenção da empresa. Se a emissão ocorreu após a data final, acumula pontuação de fraude.

Fonte: Adaptado de Lins Neto (2021)

Para validar o desempenho da ferramenta, o pesquisador utilizou um conjunto de dados com 3.080 NFS-e do setor da Construção Civil, emitidas entre os exercícios de 2010 e 2013. Na avaliação da tarefa de classificação, a ferramenta obteve 0,86 de Acurácia. Já na tarefa de extração de informações, obteve 0,90 de Precisão.

Um dos principais problemas relatado pelo autor foi a dificuldade de extração das informações quando escritas fora dos padrões pré-estabelecidos nas regras, como por exemplo, “01 a 25/09/2015” ou “26/12/2014A 25/01/2015”. Além de comprometer o desempenho da ferramenta, isso exige constantes manutenções para adaptação das regras à realidade dos dados que estão sendo gerados, já que o usuário responsável pela emissão da NFS-e pode digitar os dados em diferentes formatos. Nesse sentido, este estudo avança ao empregar técnicas de aprendizado de máquina, permitindo maior generalização e adaptabilidade frente à variabilidade das descrições em NFS-e.

O estudo realizado por De Araujo Neto (2021), por meio de uma parceria entre a Universidade Estadual da Paraíba (UEPB) e a Secretaria da Fazenda da Paraíba (SEFAZ-PB), teve como objetivo desenvolver um classificador para identificação da Nomenclatura Comum do Mercosul (NCM) do produto a partir de sua descrição na NF-e. Para isso, ele avaliou o desempenho de quatro modelos utilizando Redes Neurais Convolucionais (CNN) e *Long Short-Term Memory* (LSTM), um tipo de Redes Neurais Recorrente (RNN). Para cada rede neural foram treinados e avaliados dois classificadores, um para cada modelo de *word embeddings* utilizado na pesquisa: W*ord2Vec* e *FastText*. A pesquisa utilizou um conjunto de dados com 6,5 milhões de NF-e.

Tabela 10 - Acurácias dos conjuntos de teste e validação.

	CNN Word2Vec	**CNN FastText**	**LSTM Word2Vec**	**LSTM FastText**
Acurácia teste	73,70%	73,50%	97,68%	97,47%
Acurácia validação	73,99%	73,79%	97,66%	97,46%

Fonte: Reprodução de De Araujo Neto (2021)

Conforme a Tabela 10, De Araujo Neto (2021) concluiu em seu estudo que os modelos baseados em redes neurais LSTM apresentaram desempenho significativamente superior aos modelos baseados em CNN, independentemente do tipo de *word embeddings* utilizado. Os modelos LSTM alcançaram resultados promissores, com o *Word2Vec* performando ligeiramente melhor do que o *FastText*. A pesquisa foi pensada no desenvolvimento, avaliação e validação dos modelos, não propondo nenhuma técnica para implantação do modelo preditivo. Diferentemente desse trabalho, voltado ao âmbito do comércio de mercadorias, o presente estudo foca no setor de serviços, propondo uma solução prática que pode ser diretamente integrada a sistemas internos da prefeitura, além de abordar o problema como uma tarefa de classificação multirrótulo.

Segundo Dias e Becker (2017), um método utilizado por empresas para recolherem menos impostos é a alteração indevida do domicílio tributário da NFS-e. A estratégia envolve a seleção de um tipo de serviço em que o recolhimento ocorre obrigatoriamente no município de prestação do serviço, mas divergente do serviço efetivamente prestado que foi descrito na discriminação da nota, que por sua vez, deveria tributar no município sede da empresa. Um dos motivos dessa prática é quando a alíquota é menor no outro município.

Para mitigar a evasão de ISS para outros municípios, os pesquisadores utilizaram as *Support Vector Machines* (SVM) para desenvolvimento de um modelo de classificação capaz de identificar NFS-e candidatas à fiscalização. O experimento fez uso de um conjunto de dados com 5.128 NFS-e do setor da Construção Civil emitidas entre janeiro e outubro de 2016 no município de Porto Alegre–RS. Os resultados obtidos nos experimentos apontaram para uma precisão entre 0,79 e 0,98 na tarefa de identificação de notas candidatas à fiscalização. Enquanto o foco dos autores é a detecção de alterações fraudulentas no domicílio tributário, o presente estudo foca na identificação precisa do código de serviço descrito na nota fiscal.

De acordo com Soares e Cunha (2020), algumas empresas declaram o recolhimento do imposto corretamente com o intuito de evitar a fiscalização, mas não o repassam ao Estado. Para os autores, esses contribuintes são categorizados como devedores contumaz, pois exercem essa prática de maneira rotineira, sistemática e de cunho criminal. No estudo foram desenvolvidos três modelos de ML para classificar contribuintes com risco de se tornarem devedores contumazes, um para cada algoritmo avaliado na pesquisa: *Logistic Regression*, *Random Forest* e *LightGBM*.

Os modelos foram treinados em uma base com 58.436 registros com dados fiscais de 20.737 contribuintes, desses 5.796 eram devedores contumazes e 14.941 contribuintes regulares. Ao final do estudo, os autores concluíram que o algoritmo *Random Forest* obteve desempenho superior em relação aos demais, com acurácia de 85,21%. Já os algoritmos *Logistic Regression* e *LightGBM* alcançaram, respectivamente, 79,74% e 81,86% na mesma métrica de avaliação. Embora os autores tenham utilizado dados fiscais, o foco do estudo deles foi a predição do comportamento do contribuinte, o que difere do presente trabalho, que se concentra especificamente na análise textual da discriminação das NFS-e.

Tabela 11 - Comparativo dos trabalhos relacionados.

Pesquisa	Técnica Utilizada	Algoritmos	Modelos	Método de Implantação
Gomes (2023)	Aprendizado de Máquina	*Linear Support Vector* *Random Forest* *Decision Tree* *Gradient Boosting* *Naïve Bayes Bernoulli* *Naïve Bayes Multinomial* *Naïve Bayes Gaussian*	Foram desenvolvidos 3 modelos de classificação para cada algoritmo. O primeiro para identificar o tipo de produto OPME. O segundo para identificar a classe do produto e o terceiro para identificar o possível procedimento SUS utilizado pelo produto.	Sistema Web com uma tela de predição composta por campo de entrada do tipo texto para informar a descrição do produto e exibir os resultados as predições.

Lins Neto (2021)	Regras explícitas	Não se aplica	Foram criadas 6 regras explícitas para extração de dados específicos da discriminação da nota. Para classificação, foram criadas 4 regras para identificar NFS-e suspeitas de fraude.	Sistema que permite importar um conjunto de notas em arquivo CSV e devolve um novo arquivo com os resultados após o processamento das extrações e classificações.
De Araujo Neto (2021)	Aprendizado de Máquina	LSTM CNN	Foram avaliados 4 classificadores utilizando as redes neurais CNN e LSTM com *Word2Vec* e *FastText*. Assim, foram gerados dois modelos para cada rede neural, um para cada *word embedding* para identificar a NCM do produto a partir de sua descrição na NF-e.	Não foi proposto. A pesquisa se manteve no âmbito do experimento.
Dias e Becker (2017)	Aprendizado de Máquina	*Support Vector Machines* (SVM)	Foi desenvolvido um modelo para classificar NFS-e que tiveram seus domicílios tributários alterados indevidamente como candidatas à fiscalização.	Não foi proposto. A pesquisa se manteve no âmbito do experimento.
Soares e Cunha (2020)	Aprendizado de Máquina	*Logistic Regression* *Random Forest* *LightGBM*	Foram desenvolvidos 3 modelos com objetivo de classificar contribuintes com risco de inadimplemento contumaz de ISS. No estudo, o *Random Forest* obteve o melhor desempenho.	Não foi proposto. A pesquisa se manteve no âmbito do experimento.
Nota Conforme	Aprendizado de Máquina	*Random Forest* *Decision Tree* *Linear Support Vector* *LightGBM* *Logistic Regression* *Gradient Boosting* *Naive Bayes Multinomial*	Foram avaliados sete algoritmos de ML. O *Random Forest* foi modelo que apresentou o melhor resultado na classificação de serviços descritos na discriminação da NFS-e. Por esse motivo, ele foi escolhido para compor o Nota Conforme.	Implantação por meio de uma API genérica com o modelo encapsulado para permitir a integração com sistemas internos dos órgãos controladores. Por utilizar campos dinâmicos, essa API pode ser facilmente implementada em diferentes órgãos.

Fonte: Elaborado pelo autor

A Tabela 11 apresenta um resumo dos trabalhos relacionados a esta pesquisa, evidenciando suas técnicas, algoritmos e método de implantação proposto. Esses trabalhos

contribuíram efetivamente para o desenvolvimento desta pesquisa, permitindo avançar conhecimento por meio de três contribuições importantes.

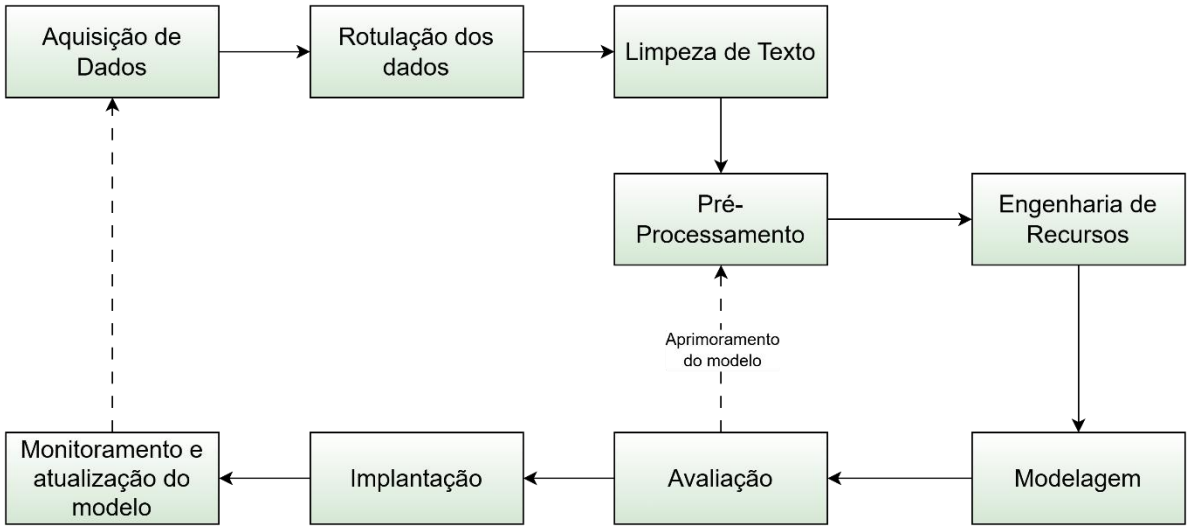
- Desenvolvimento de um modelo de classificação multirrótulo capaz de identificar todos os códigos de serviços descritos na discriminação da NFS-e.
- A possibilidade de identificar discriminações genéricas, quando não descrevem um serviço prestado de forma clara, conforme estabelecido no artigo 2º do Decreto n.º 24.093/2008 (RECIFE, 2008b).
- Desenvolvimento de uma API para implantação do modelo treinado em ambiente de produção, permitindo a integração com outros sistemas dos órgãos de controle.

4. MATERIAIS E MÉTODOS

O desenvolvimento de sistemas com Processamento de Linguagem Natural (PLN) envolve diversas etapas, cujo objetivo é preparar os dados, construir e avaliar o modelo. O conjunto organizado dessas etapas forma o processo conhecido como *pipeline* de dados (VAJJALA et al., 2020, p. 37). Segundo os autores, os principais componentes de um *pipeline* genérico de PLN são: aquisição de dados, limpeza de texto, pré-processamento, engenharia de recursos, modelagem, avaliação, implantação e monitoramento e atualização do modelo.

Por se tratar de uma solução baseada em aprendizado supervisionado, é necessário fornecer ao algoritmo dados de treinamento que já incluem rótulos em suas instâncias (GÉRON, 2021, p. 8). Diante disso, e da ausência de uma base de dados previamente rotulada, o *pipeline* proposto por Vajjala et al. (2020, p. 38) foi adaptado para o contexto desta pesquisa, com a inclusão de um novo componente para rotulação dos dados, executado após a aquisição deles, conforme apresentado na Figura 3.

Figura 3 - Pipeline de um projeto de PLN.



Fonte: Adaptado de Vajjala et al. (2020, p. 38)

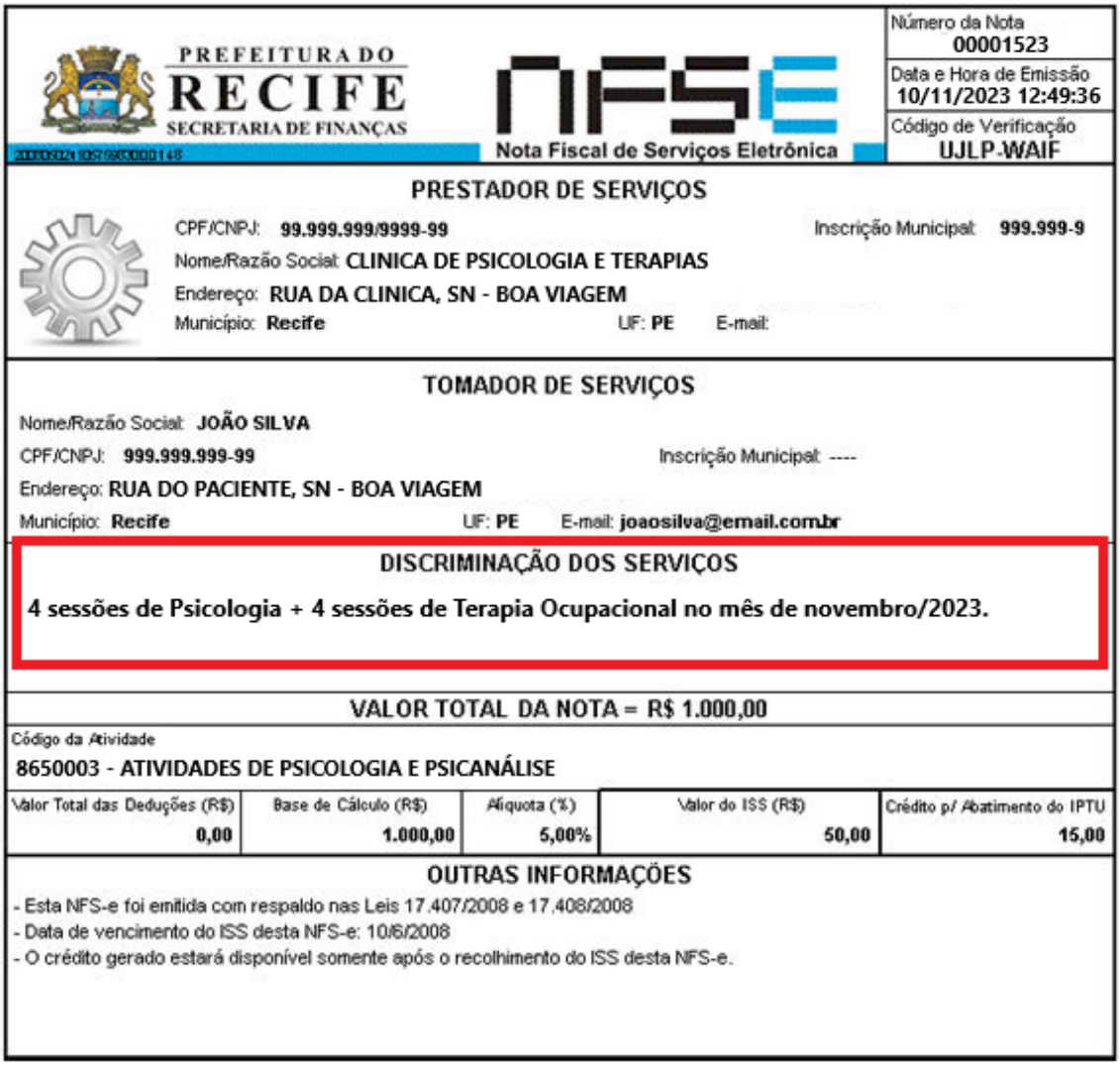
Assim, esta seção apresentará os procedimentos metodológicos e as técnicas utilizadas para desenvolvimento do modelo de classificação e a viabilização de sua implantação em ambiente de produção, detalhando as tarefas executadas em cada etapa do *pipeline* da solução desenvolvimento, o que permite sua reprodutibilidade em trabalhos futuros, para aprimorá-lo.

4.1. Aquisição dos dados

A coleta de dados é o primeiro passo para desenvolvimento de qualquer sistema baseado em PLN (VAJJALA et al., 2020, p. 38). Em um conjunto de dados, cada objeto, conhecido também como instância, representa uma ocorrência dos dados e cada atributo corresponde a uma propriedade do objeto. Com isso, os objetos são caracterizados por um conjunto de atributos (FACELI et al., 2023, p. 10).

Apesar das tabelas que armazenam os dados das Notas Fiscais de Serviços Eletrônicas (NFS-e) possuírem dezenas de atributos, para criação da base de treinamento e teste foi utilizado como atributo preditor somente o campo discriminação da NFS-e, conforme destacado na Figura 4. Dessa forma, os demais campos foram descartados para construção do modelo.

Figura 4 - Visão do campo Discriminação da NFS-e.


PREFEITURA DO RECIFE
SECRETARIA DE FINANÇAS

NFS-e
Nota Fiscal de Serviços Eletrônica

Número da Nota 00001523
Data e Hora de Emissão 10/11/2023 12:49:36
Código de Verificação UJLP-WAIF

PRESTADOR DE SERVIÇOS
CPF/CNPJ: 99.999.999/9999-99 Inscrição Municipal: 999.999-9
Nome/Razão Social: CLINICA DE PSICOLOGIA E TERAPIAS
Endereço: RUA DA CLINICA, SN - BOA VIAGEM
Município: Recife UF: PE E-mail:

TOMADOR DE SERVIÇOS
Nome/Razão Social: JOÃO SILVA
CPF/CNPJ: 999.999.999-99 Inscrição Municipal: ----
Endereço: RUA DO PACIENTE, SN - BOA VIAGEM
Município: Recife UF: PE E-mail: joaosilva@email.com.br

DISCRIMINAÇÃO DOS SERVIÇOS
4 sessões de Psicologia + 4 sessões de Terapia Ocupacional no mês de novembro/2023.

VALOR TOTAL DA NOTA = R$ 1.000,00

Código da Atividade
8650003 - ATIVIDADES DE PSICOLOGIA E PSICANÁLISE

Valor Total das Deduções (R$)	Base de Cálculo (R$)	Alíquota (%)	Valor do ISS (R$)	Crédito p/ Abatimento do IPTU
0,00	1.000,00	5,00%	50,00	15,00

OUTRAS INFORMAÇÕES
- Esta NFS-e foi emitida com respaldo nas Leis 17.407/2008 e 17.408/2008
- Data de vencimento do ISS desta NFS-e: 10/6/2008
- O crédito gerado estará disponível somente após o recolhimento do ISS desta NFS-e.

Fonte: Adaptado de Recife (2024a)

As NFS-e extraídas foram armazenadas em uma base local do PostgreSQL, em tabela representada na Figura 11. O campo *discriminacao* (Item 1), trata-se do campo de discriminação da NFS-e, originado da base de dados. Já os campos apresentados no Item 2 foram utilizados durante o processo de rotulação para identificar os serviços descritos em cada uma das NFS-e. E no campo *is_1* (Item 3), foram registradas as NFS-e com discriminações genéricas. Esse processo de rotulação está descrito na sessão 4.2.

Tabela 12 - Campos para criação da base de treinamento.

ITEM	NOME	TIPO	DESCRIÇÃO
1	discriminacao	Alfanumérico	Discriminação da NFS-e. Atributo preditor.
2	is_0401; is_0402 ... is_0421	Inteiro	Identifica se possui serviços correspondentes aos códigos dos serviços na discriminação da NFS-e (1 para sim; 0 para não). Atributos classes
3	is_1	Inteiro	Identifica se a discriminação é genérica ou não pertence a nenhum código de serviço da lista (1 para sim; 0 para não). Atributo classe.

Fonte: Elaborado pelo autor

Para esta pesquisa, foram coletadas no *Data Warehouse* (DW) aproximadamente 6,56 milhões de NFS-e emitidas entre 01/01/2022 e 30/06/2024. As notas extraídas estão distribuídas nas 21 atividades de prestação de serviços do setor de saúde apresentadas na Tabela 13. O segmento foi escolhido inicialmente devido à sua relevância apontada por auditores, especialmente em relação ao volume de notas e à ocorrência de inconsistências. A inclusão de novos setores está prevista para versões futuras. Finalizada a coleta de dados, eles foram disponibilizados para os devidos tratamentos executados nas próximas etapas.

Tabela 13 - Tipos de Serviços do setor de saúde, contemplados na pesquisa.

CÓDIGO	TIPO DE SERVIÇO - SETOR SAÚDE
04.01	Medicina e biomedicina.
04.02	Análises clínicas, patologia, eletricidade médica, radioterapia, quimioterapia, ultrassonografia, ressonância magnética, radiologia, tomografia e congêneres.
04.03	Hospitais, clínicas, laboratórios, sanatórios, manicômios, casas de saúde, prontos socorros, ambulatórios e congêneres.
04.04	Instrumentação cirúrgica.
04.05	Acupuntura.
04.06	Enfermagem, inclusive serviços auxiliares.
04.07	Serviços farmacêuticos
04.08	Terapia ocupacional, fisioterapia e fonoaudiologia.
04.09	Terapias de qualquer, espécie destinadas ao tratamento físico, orgânico e mental.

04.10	Nutrição.
04.11	Obstetrícia.
04.12	Odontologia.
04.13	Ortóptica.
04.14	Próteses sob encomenda.
04.15	Psicanálise.
04.16	Psicologia.
04.17	Casas de repouso e de recuperação, creches, asilos e congêneres.
04.18	Inseminação artificial, fertilização in vitro e congêneres.
04.19	Bancos de sangue, leite, pele, olhos, óvulos, sêmen e congêneres.
04.20	Coleta de sangue, leite, tecidos, sêmen, órgãos e materiais biológicos de qualquer espécie.
04.21	Unidade de atendimento, assistência ou tratamento móvel e congêneres.

Fonte: Recife (2024a)

4.2. Rotulação da amostra para treinamento

De acordo com Kang et al. (2020, p. 22), a rotulação de dados é um grande desafio em projetos de aprendizado supervisionado. Diante da inexistência de uma base de dados previamente rotulada para o treinamento do modelo, tornou-se necessário realizar uma análise detalhada das NFS-e extraídas, a fim de identificar e rotular os serviços prestados. Esta etapa foi realizada conforme as orientações e as regras de negócios definidas pelos auditores.

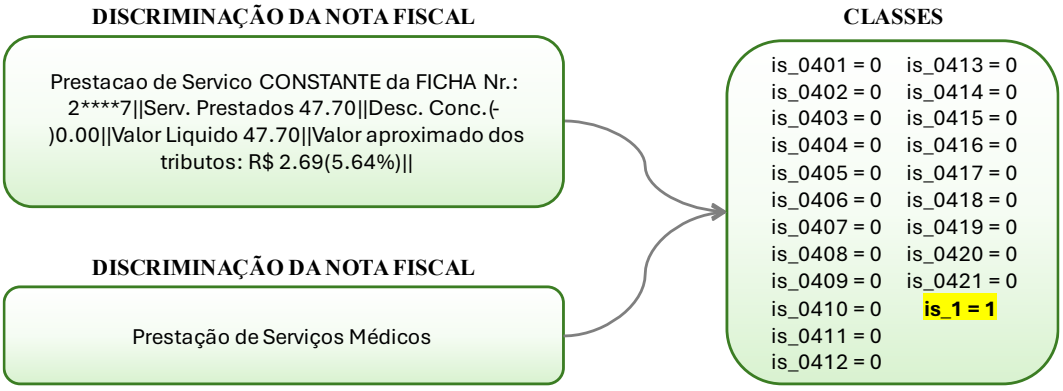
O campo de discriminação da NFS-e exerce um papel central nesse processo, pois se trata de um campo de texto livre onde o prestador deve descrever detalhadamente os serviços prestados (RECIFE, 2024b, p. 42). Dessa forma, as NFS-e foram rotuladas com base na análise das descrições textuais. Esse procedimento foi realizado por meio da combinação entre inspeção manual e o uso de comandos SQL, elaborados a partir de padrões identificados ao longo da análise. Essa abordagem permitiu automatizar parte do processo de rotulação e ampliar a quantidade de NFS-e rotuladas.

A rotulação resultou em 22 classes distintas: 21 correspondentes aos códigos de serviços do setor de saúde e uma classe adicional destinada às notas com descrições genéricas, nas quais não foi possível identificar um serviço específico. A criação de uma classe adicional tornou-se necessária devido à presença de descrições que não identificam objetivamente o serviço prestado, configurando um erro na emissão da NFS-e, além de contrariar a legislação vigente.

Diante disso, os auditores da Secretaria de Finanças da Prefeitura Municipal do Recife–PE destacaram a importância de identificar esses casos. Para esse fim, foi definida a classe

complementar, denominada “is_1”, conforme apresentado no item 3 da Tabela 12. Assim, todas as NFS-e que apresentaram essas características foram associadas nessa nova categoria, conforme ilustrado na Figura 5.

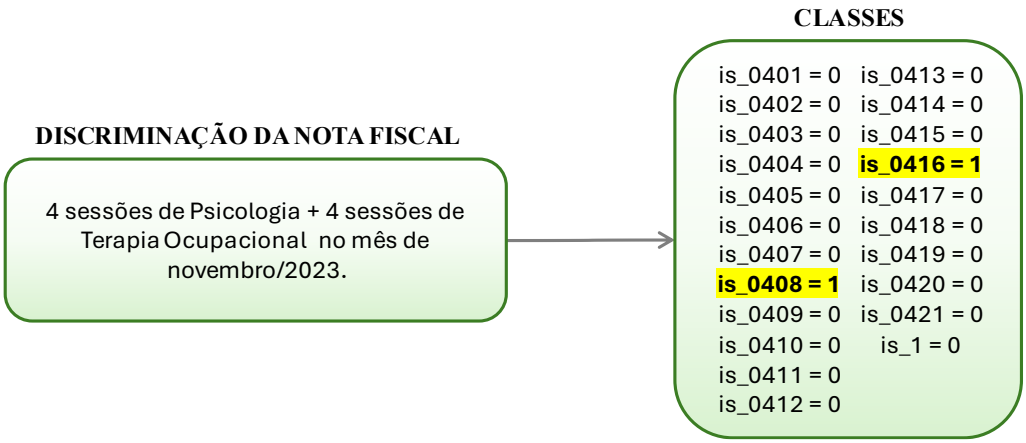
Figura 5 – Exemplos de discriminações rotuladas como genéricas.



Fonte: Elaborado pelo autor

Além disso, *“o prestador de serviços deverá emitir uma Nota Fiscal para cada serviço prestado, sendo vedada a emissão de uma Nota Fiscal que englobe serviços enquadrados em mais de um código de atividade”* (RECIFE, 2024b, p. 41). No entanto, na prática, é comum a emissão de NFS-e, que descrevem múltiplos serviços em um único documento. Um exemplo disso é apresentado na Figura 6, em que a NFS-e inclui, em sua descrição, tanto o serviço de psicologia (código 0416) quanto o de terapia ocupacional (código 0408). Diante desse contexto, uma mesma nota fiscal pode ser associada a múltiplas classes de serviço, caracterizando um cenário de classificação multirrótulo (FACELI et al., 2023, p. 264).

Figura 6 - Exemplos de notas com mais de um rótulo.



Fonte: Elaborado pelo autor

A disponibilização de dados relevantes e representativos é fundamental para o desenvolvimento de modelos de aprendizado de máquina (ML) (KALINOWSKI et al., 2023, p. 125). A capacidade de generalização está intrinsicamente relacionada à utilização de dados de treinamento com boas representações para novos casos, pois o modelo dificilmente fará predições exatas com a utilização de conjuntos de dados de treinamento não representativos (GÉRON, 2021, p. 20). Assim, para garantir ampla representatividade dos dados coletados, foram rotuladas 5,69 milhões de NFS-e, aproximadamente 86,73% da base de dados extraída. Concluída essa etapa, os dados ficaram prontos para o pré-processamento.

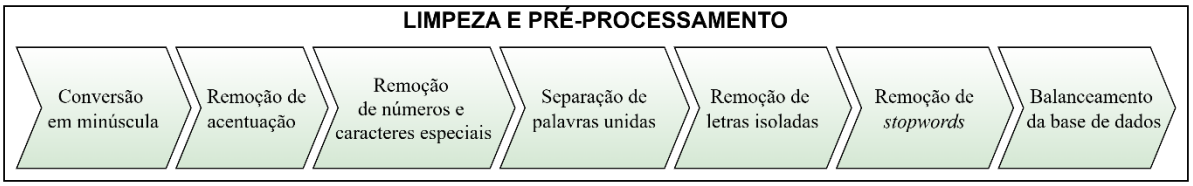
4.3. Limpeza de texto e pré-processamento dos dados

A utilização de dados de alta qualidade é essencial em projetos de ML para garantir que o modelo produza previsões assertivas (KALINOWSKI et al., 2023, p. 121). Entretanto, ao trabalharmos com dados, é comum que estes apresentem lacunas, ruídos e inconsistências. Esses problemas podem afetar negativamente o desempenho dos modelos de ML. Por meio das técnicas de pré-processamento é possível eliminar ou mitigar esses problemas, melhorando a qualidade dos dados. Além disso, o pré-processamento pode ser útil para tornar os dados mais adequados para um determinado algoritmo (FACELI et al., 2023, p. 28).

De acordo com Kalinowski et al. (2023, p. 241, 255), o pré-processamento é uma etapa fundamental, pois é nela que os dados são preparados para obter resultados melhores nas etapas seguintes. Ainda segundo os autores, o número de características de um conjunto de dados define a sua dimensionalidade. Muitas dessas características podem ser irrelevantes para o desempenho do modelo de classificação. Nesse contexto, efetuar a redução da dimensionalidade, preservando as propriedades mais relevantes do conjunto de dados é essencial, pois muitos algoritmos de ML tendem a apresentar melhores resultados quando aplicados a conjuntos de dados com dimensionalidades reduzida.

Dessa forma, considerando se tratar, sobretudo, de um campo de texto livre, o pré-processamento desempenhou um papel fundamental na redução de ruídos, eliminação de características irrelevantes e redução da dimensionalidade, com objetivo de melhorar a qualidade dos resultados da pesquisa. Esse processo foi realizado em sete etapas, conforme apresentado na Figura 7. Em vista disso, esta sessão se destina a detalhar essas etapas para melhor compreensão de todo processo.

Figura 7 - Etapas da limpeza e pré-processamento dos dados.

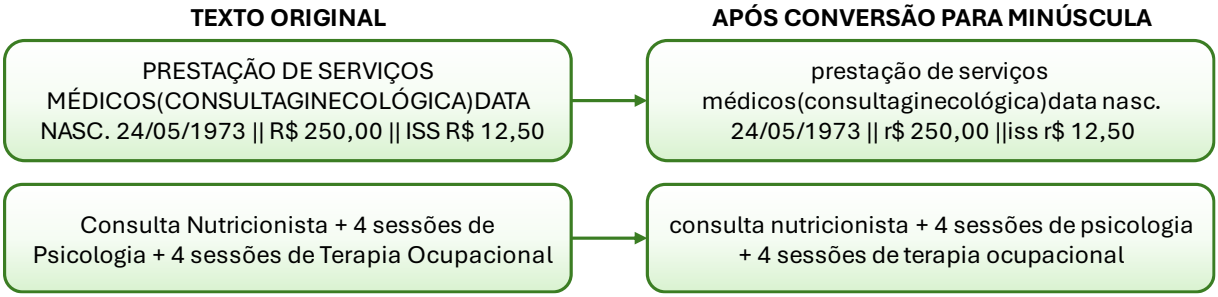


Fonte: Elaborado pelo autor

4.3.1. Conversão do texto em minúscula

As mesmas palavras quando escritas somente com diferença entre utilização de letras maiúsculas e minúsculas são tratadas como distintas pelo algoritmo. Para obtenção de melhor precisão, é fundamental que as letras do documento sejam convertidas para suas respectivas formas minúsculas (UYSAL; GUNAL, 2014, p. 106, 110).

Figura 8 - Resultado da conversão para minúsculo.



Fonte: Elaborado pelo autor

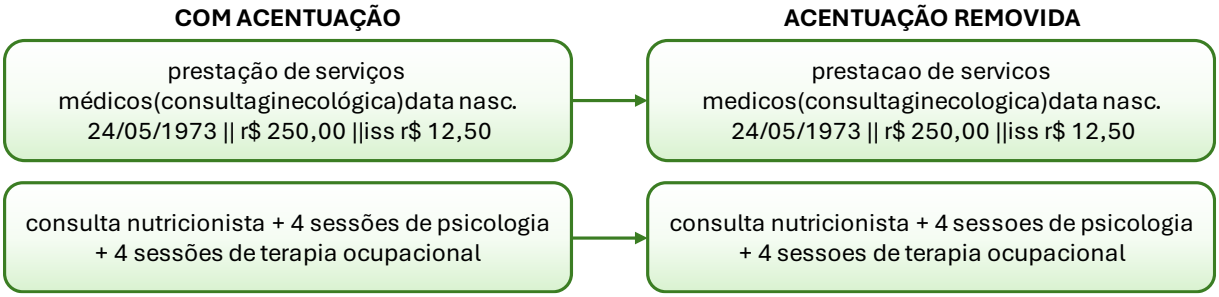
Logo, a primeira etapa foi responsável pela conversão de todo o conteúdo das discriminações das NFS-e do conjunto de dados em minúsculo. Para isso, foi utilizado o método *lower()* da biblioteca Pandas, que ao receber o conteúdo de uma coluna do tipo texto, efetua sua conversão e retorna uma cópia com todas as letras em minúsculas (PANDAS DEVELOPMENT TEAM, 2024).

4.3.2. Remoção de acentuação

O problema da acentuação é semelhante à variação entre letras maiúsculas e minúsculas observado na etapa anterior. O processo de conversão também ocorre de forma semelhante, atuando ao nível de caracteres. Os caracteres acentuados são substituídos pelos mesmos caracteres, porém sem os acentos (CHAI, 2023, p. 519).

Para essa tarefa, utilizou-se a função *normalize*() do módulo Unicodedata do Python. Que ao utilizá-la com argumento NFKD (*Normalization Form KC, Compatibility Composition*), efetua decomposição dos caracteres em seus componentes básicos (PYTHON SOFTWARE FOUNDATION, 2024).

Figura 9 - Substituição de letras acentuadas por sua correspondência sem acentuação.



Fonte: Elaborado pelo autor

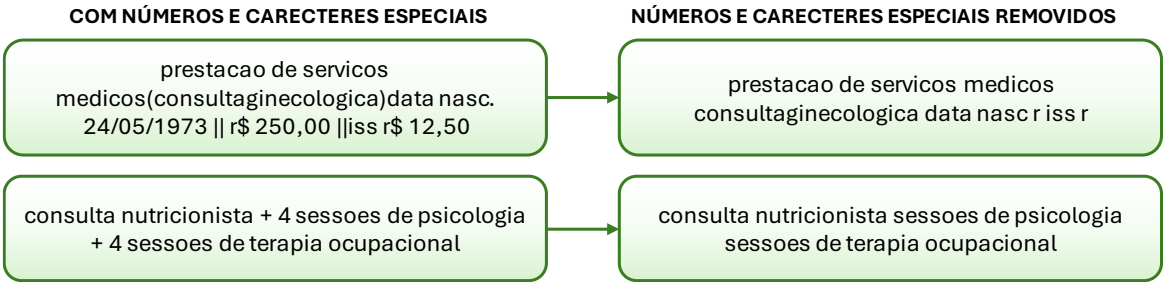
Portanto, a segunda etapa consistiu na substituição das letras acentuadas por suas respectivas formas sem acento, conforme ilustrado na Figura 9. Com isso, independentemente da forma de escrita utilizada pelo usuário ao emitir as NFS-e, a palavra terá a mesma importância para o modelo de classificação.

4.3.3. Remoção de números e caracteres especiais

Durante a rotulação das NFS-e foi identificada a presença de diversos números e caracteres especiais para representar os valores dos serviços e de impostos, contatos, número de protocolos, documentos, datas e diversas outras informações relevantes para os prestadores ou tomadores dos serviços. Apesar de permitida a sua inclusão na descrição da NFS-e, nenhuma dessas informações exercem importância para a identificação de um serviço de saúde.

Nesse sentido, Géron (2021), explica que um sistema não será capaz de aprender o suficiente para ter um bom desempenho se o conjunto de treinamento estiver poluído com características irrelevantes. De acordo com Vajjala et al. (2020), é comum incluir uma etapa para remoção de números e pontuações em problemas de classificação com PLN. Desse modo, a terceira etapa removeu todos os números e caracteres especiais, conforme apresentado na Figura 10.

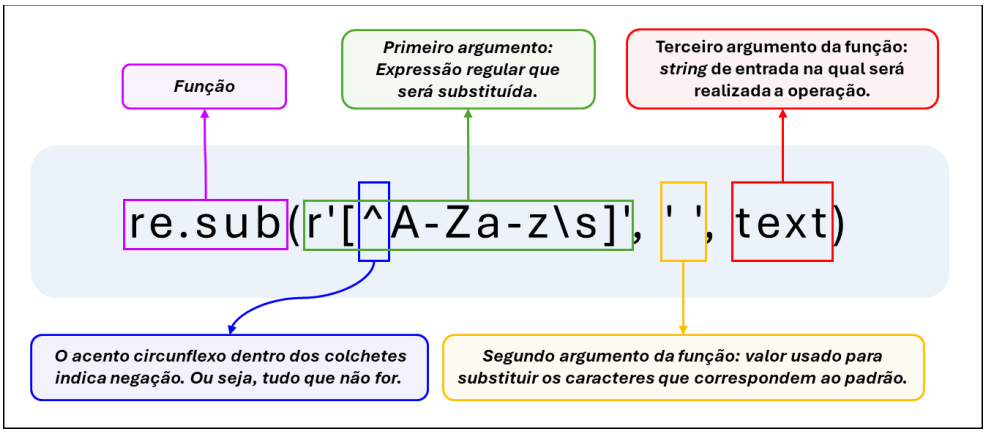
Figura 10 - Exemplo de remoção de número, especiais e espaços excessivos.



Fonte: Elaborado pelo autor

A remoção dos números e caracteres especiais foi executada através da biblioteca de operações com expressões regulares do Python, conhecida como *re*. A função *sub* da biblioteca substitui parte de uma *string* com base em um padrão especificado em outra *string* fornecida como parâmetro de entrada (PYTHON SOFTWARE FOUNDATION, 2024).

Figura 11 - Detalhamento da função para remover números e caracteres especiais.



Fonte: Elaborado pelo autor

Conforme a Figura 11, o primeiro argumento define a expressão regular que será substituída. A expressão "A-Za-z" busca qualquer letra, seja ela maiúscula ou minúscula, enquanto "\s" procura por qualquer espaço em branco. O acento circunflexo nos colchetes indica uma negação da expressão. Ou seja, essa combinação define que qualquer caractere que não seja uma letra ou espaço será substituído. O segundo argumento da função define o valor utilizado para substituir os caracteres que correspondem ao padrão. O terceiro argumento é a *string* na qual as substituições serão realizadas.

Figura 12 - Função para remover espaços excessivos.

```
text = " ".join(text.split())
```

Fonte: Elaborado pelo autor

Adicionalmente, essa etapa também removeu todos os espaços excessivos. Para isso, conforme a Figura 12, foi utilizado a combinação dos métodos *split()* que retorna uma lista de palavras, separando-as com base em espaços em branco. E o método *join()*, utilizado para unir os elementos de uma lista em única *string*, permitindo a definição de um separador entre esses elementos (PYTHON SOFTWARE FOUNDATION, 2024). Nessa etapa, o separador é um único espaço, inserido entre as palavras.

4.3.4. Identificação e separação de palavras unidas

Além disso, foram localizadas ocorrências de palavras unidas. Essas inconsistências poderiam gerar resultados significativamente comprometidos, pois ao fazer a transformação em *tokens* as palavras unidas resultariam em somente um *token* com pouca ou nenhuma representatividade no documento. Considerando o conjunto de dados de treinamento apresentado na Tabela 4, ao submeter os termos dos documentos apresentados na Tabela 14 para avaliar suas respectivas importâncias para identificar um documento semelhante ao D3, que se refere a uma consulta ginecológica, foi atribuído o valor zero de *Term Frequency-Inverse Document Frequency* (TF-IDF) quando as palavras foram mantidas unidas. Já ao realizar a separação das palavras, foram gerados dois *tokens* com relevância para o documento.

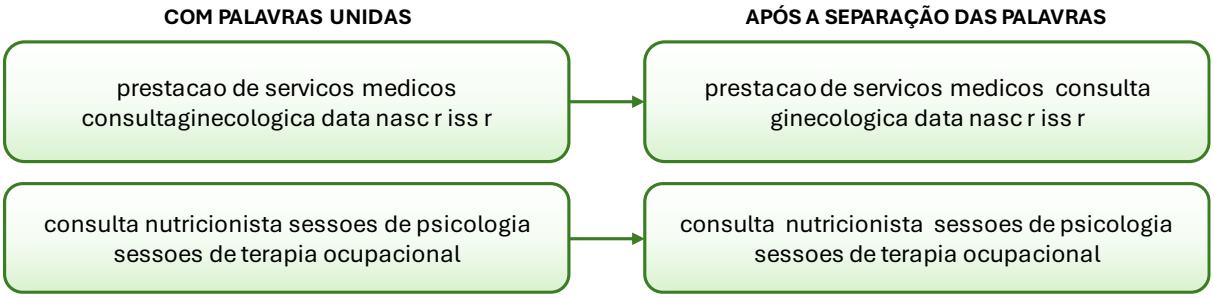
Tabela 14 - Comparativo de importância dos termos unidades e pré-processados.

DOCUMENTO	TOKEN	TF-IDF
CONSULTAGINECOLOGICA	CONSULTAGINECOLOGICA	0
CONSULTA GINECOLOGICA	CONSULTA	0.0959
	GINECOLOGICA	0.4621

Fonte: Elaborado pelo autor

Para mitigar esse problema, a quarta etapa do pré-processamento se concentrou na busca por palavras unidas em todo o conjunto de dados, seguida de sua separação. Isso foi possível através da criação de um dicionário de palavras comumente utilizadas na prestação de serviços da área de saúde que possuem relevância para identificação de especialidades, exames e procedimentos médicos. O dicionário é composto por 634 palavras, disponibilizadas no repositório do sistema no GitHub. A função em *Python* busca a ocorrência dessas palavras, verificando se estão antecedendo ou na sequência de outra palavra e faz a separação incluindo um espaço em branco entre elas, transformando em duas palavras distintas, conforme ilustrado na Figura 13.

Figura 13 - Exemplos após separação das palavras unidas.

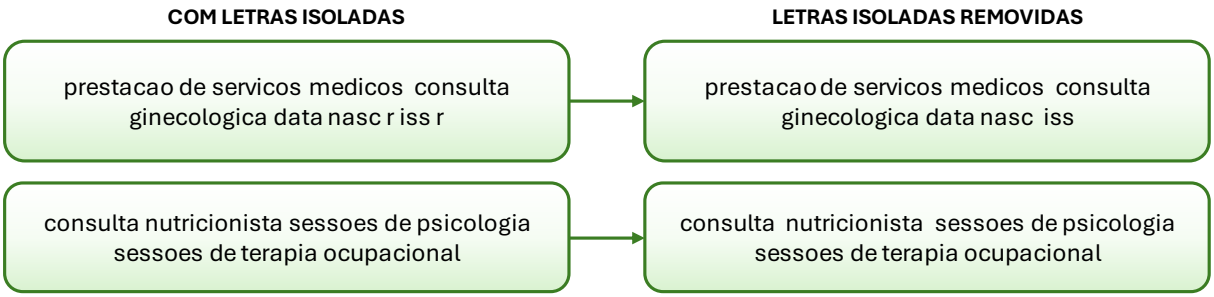


Fonte: Elaborado pelo autor

4.3.5. Remoção de letras isoladas

Durante o pré-processamento foram identificadas diversas letras isoladas no conjunto de dados. Esses ruídos, além de ter sido obtidos direto da fonte de dados de origem, também foram resultados dos tratamentos realizados nas etapas anteriores. Portanto, a quinta etapa efetuou mais uma camada de limpeza do conjunto de dados, eliminando todas as letras isoladas das discriminações das NFS-e, conforme resultado ilustrado na Figura 14.

Figura 14 - Exemplos após a remoção de letras isoladas.



Fonte: Elaborado pelo autor

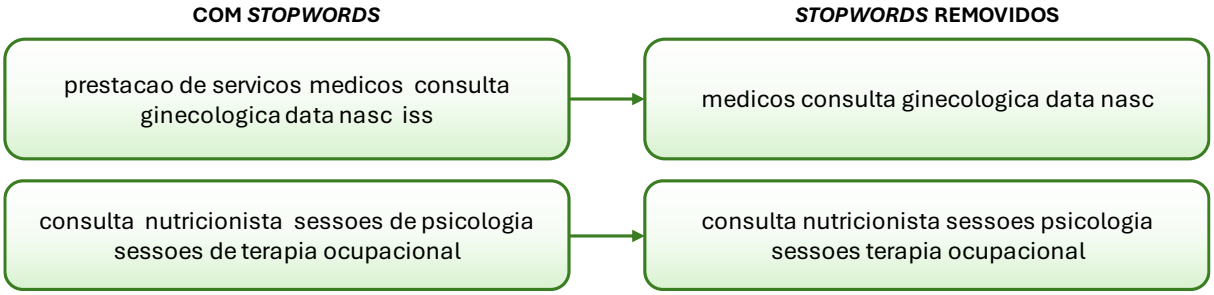
4.3.6. Remoção de *stopwords*

Os *Stopwords* são palavras que aparecem com muita frequência em qualquer tipo de documento, mas sua presença não fornece informações sobre o conteúdo específico do texto. Ou seja, são palavras consideradas irrelevantes que é recomendável remover durante as tarefas de PLN. São exemplos de *stopwords* os artigos, preposições, pronomes (FERILLI; ESPOSITO; GRIECO, 2014, p. 118-119).

As tarefas de remoção de *stopwords* são normalmente realizadas a partir de lista pré-definida de palavras-chave. Entretanto, outros termos podem ser adicionados à lista, pois eles

podem ser considerados insignificantes em domínios específicos (FERILLI; ESPOSITO; GRIECO, 2014, p. 118-119). Com isso, além de utilizar termos padrões da língua portuguesa disponibilizados pela biblioteca *Natural Language Toolkit* (NLTK) do *Python*, foram adicionadas 71 palavras (disponíveis no repositório do sistema no GitHub) frequentemente utilizados nas emissões das notas, mas que não possuem nenhuma relevância para identificação do serviço prestado.

Figura 15 - Exemplos após a remoção de *stopwords*.



Fonte: Elaborado pelo autor

Portanto, a sexta etapa se concentrou na remoção dos *stopwords*, conforme exemplos apresentados na Figura 15.

4.3.7. Balanceamento da base de dados

A partir da análise da distribuição das classes no conjunto de dados original, observou-se que determinados serviços são registrados com maior frequência do que outros, causando um forte desbalanceamento. Classes majoritárias concentravam uma grande proporção das amostras como, por exemplo, as classes is_0401 e is_0402 representavam, respectivamente, 20,40% e 24,97% do total. Em contrapartida, diversas classes minoritárias, como is_0413, is_0417 e is_0420, possuíam menos de 0,1% de representatividade., conforme apresentado na Tabela 15.

Tabela 15 – Distribuição das classes na base de treinamento

CLASSE	PRÉ-BALANCEAMENTO		PÓS-BALANCEAMENTO	
	QUANTIDADE	PROPORÇÃO	QUANTIDADE	PROPORÇÃO
is_0401	1.160.728	20,40%	127.115	5,86%
is_0402	1.420.819	24,97%	175.828	8,11%
is_0403	742.779	13,05%	246.189	11,35%
is_0404	2.396	0,04%	100.332	4,63%
is_0405	4.881	0,09%	100.013	4,61%

is_0406	2.666	0,05%	100.022	4,61%
is_0407	918.688	16,15%	100.052	4,61%
is_0408	297.622	5,23%	186.798	8,61%
is_0409	29.845	0,52%	127.562	5,88%
is_0410	43.845	0,77%	108.963	5,02%
is_0411	14.982	0,26%	103.683	4,78%
is_0412	651.774	11,45%	137.617	6,35%
is_0413	1.184	0,02%	100.000	4,61%
is_0414	20.636	0,36%	100.973	4,66%
is_0415	24.098	0,42%	119.981	5,53%
is_0416	227.550	4,00%	179.462	8,27%
is_0417	5.007	0,09%	100.000	4,61%
is_0418	22.093	0,39%	102.551	4,73%
is_0419	8.570	0,15%	100.963	4,66%
is_0420	5.383	0,09%	100.006	4,61%
is_0421	15.112	0,27%	100.157	4,62%
is_1	514.224	9,04%	103.963	4,79%

Fonte: Elaborado pelo autor

Um conjunto de dados balanceado é essencial para evitar que o modelo de classificação se torne tendencioso. Algumas técnicas como coleta de novos dados, reamostragem (subamostragem ou superamostragem) ou ajuste de pesos podem ser aplicadas para mitigar o desequilíbrio (VAJJALA et al., 2020, p. 156). Neste trabalho, foi definido um critério mínimo de 100.000 registros por classe. Com base nesse critério, aplicou-se um balanceamento híbrido, que combinou duas estratégias complementares. Inicialmente, foi aplicada a subamostragem das classes majoritárias com número de exemplos superior ao mínimo estabelecido, por meio da função **random.choice** da biblioteca Numpy. Em seguida, foi realizada a superamostragem com reposição das classes minoritárias, através da função **utils.resample** da biblioteca Scikit-learn, para que todas alcançassem a quantidade mínima requerida.

Como resultado, a base balanceada passou a conter cerca de 2,1 milhões de amostras de NFS-e, apresentando uma distribuição de classes significativamente mais uniforme. As classes que antes dominavam a base, como is_0402, tiveram sua quantidade de amostras reduzida (de 1.420.819 para 175.828), enquanto classes pouco representadas, como is_0413 (de 1.184 para 100.000), foram aumentadas artificialmente. Essa abordagem permitiu que todas as classes tivessem uma representação mais equilibrada, com proporções variando entre aproximadamente 4,6% e 11,35%, favorecendo uma aprendizagem mais justa e eficaz pelo modelo.

4.4. Engenharia de recursos

De acordo com Vajjala et al. (2020, p. 62), engenharia de recursos é o processo de transformação dos dados brutos em formato processável por uma máquina. No contexto deste estudo, o objetivo dessa etapa é transformar as características dos textos em vetores numéricos, de modo que os algoritmos de ML consigam interpretá-los. Esse processo foi realizado no atributo de discriminação da NFS-e, utilizando a técnica TF-IDF.

A escolha dessa técnica foi motivada pela sua capacidade de quantificar a relevância de cada palavra no contexto do conjunto de documentos (VAJJALA et al., 2020, p. 90). A técnica foi aplicada utilizando *uni-gramas*, conforme o padrão do método adotado. Essa abordagem implica que cada termo do vocabulário é tratado de forma isolada, desconsiderando combinações sequenciais de palavras (SCIKIT-LEARN DEVELOPER, 2024).

Para reduzir a dimensionalidade e volume dos dados, foi definido ao parâmetro *max_features* o valor *10000*. Esse parâmetro determina a quantidade máxima de palavras a serem mantidas na matriz de características, selecionando-as com base na sua relevância para o conjunto de documentos analisados. A ausência desse parâmetro indica que todos os termos serão utilizados (SCIKIT-LEARN DEVELOPER, 2024). Assim, a utilização do parâmetro contribui significativamente para melhoria da eficiência computacional, pois o modelo de classificação se concentrará nos termos mais relevantes, resultando, por consequência, em um processo de treinamento mais rápido.

Além disso, ao utilizar o TF-IDF através da biblioteca *Scikit-learn*, não é necessário realizar a conversão dos documentos em *tokens* de forma antecipada, pois esse processo é incorporado diretamente em seu fluxo de trabalho (SCIKIT-LEARN DEVELOPER, 2024).

4.5. Modelagem dos classificadores

Após a coleta, o tratamento e a transformação dos dados em um formato adequado para processamento por algoritmos de aprendizado de máquina, a próxima etapa consiste no desenvolvimento de uma solução aplicável ao problema proposto (VAJJALA et al., 2020, p. 62). Para isso, o primeiro passo é a escolha do algoritmo mais apropriado. De acordo com Faceli et al. (2023, p. 50), existem variações de desempenho nos diferentes tipos de algoritmos. Dessa forma, a escolha de um algoritmo adequado e eficiente para resolver o problema em questão é fundamental para alcançar bons resultados.

Para este trabalho, foram selecionados sete algoritmos de classificação amplamente utilizados na literatura e aplicados em estudos similares discutidos no capítulo anterior: *Random Forest, Decision Tree, Linear Support Vector, LightGBM, Logistic Regression, Gradient Boosting* e *Naive Bayes Multinomial*. A escolha desses algoritmos busca abranger diferentes abordagens, desde métodos lineares até *ensembles*, de modo a permitir uma comparação abrangente de desempenho. Cada algoritmo possui características distintas em termos de complexidade, interpretabilidade e capacidade de generalização, o que permite avaliar como cada um se comporta diante dos desafios impostos pelos dados textuais oriundos das descrições de serviços nas NFS-e.

Os algoritmos foram implementados por meio da biblioteca *Scikit-learn*, mantendo os hiperparâmetros em seus valores padrão. Essa padronização pretende assegurar uma base justa para a comparação inicial entre os modelos, evitando vieses decorrentes de ajustes específicos que poderiam favorecer indevidamente em relação aos demais. Os modelos foram treinados utilizando a técnica de validação cruzada *k-fold* com 5 partes.

A validação cruzada é uma técnica de avaliação de desempenho que consiste em dividir aleatoriamente o conjunto de dados em *k* subconjuntos, denominados de *folds*. Em cada iteração, um dos *folds* é utilizado como conjunto de teste, enquanto os *k-1* restantes são utilizados para treinamento do modelo. Esse processo é repetido *k* vezes e o desempenho final é calculado pela média das avaliações obtidas em cada subconjunto de teste (FACELI et al., 2023, p. 151-152). Essa técnica possibilita uma estimativa mais precisa da acurácia do modelo ao utilizar todo o conjunto de dados para teste e treinamento, em contraste com o método *holdout*, que emprega somente uma partição fixa do conjunto de dados (KALINOWSKI et al., 2023, p. 252).

Tabela 16 - Especificações do computador utilizado para o treinamento dos modelos.

CARACTERÍSTICA	VALOR
Processador	12th Gen Intel(R) Core(TM) i7-1260P 2.10 GHz
Núcleos	12
Processadores lógicos	16
Memória RAM	16GB
GPU	Intel(R) Iris(R) Xe Graphics
Memória da GPU compartilhada	7,8GB
Sistema Operacional	Windows 11 Pro

Fonte: Elaborado pelo autor

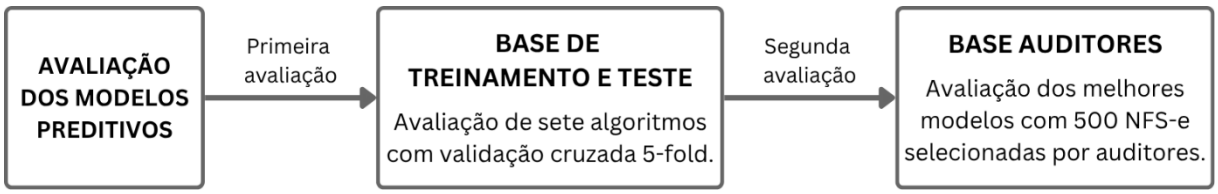
Todas as etapas foram realizadas utilizando os recursos computacionais apresentados na Tabela 16. Por fim, os modelos passaram por um processo de avaliação, utilizando métricas de desempenho que serão discutidas na próxima seção.

4.6. Avaliação dos modelos

A avaliação de desempenho é uma etapa essencial para medir a qualidade do modelo desenvolvido. O método mais comum é a avaliação realizada a partir de dados não vistos pelo modelo (VAJJALA et al., 2020, p. 68). Com isso, é possível mensurar sua capacidade de classificar corretamente os serviços em NFS-e desconhecidas para assegurar sua habilidade em generalização.

Com o intuito de aferir a eficácia e a confiabilidade dos modelos preditivos desenvolvidos no âmbito do Nota Conforme, bem como subsidiar a escolha do modelo com melhor desempenho, foram conduzidos dois procedimentos de validação distintos e independentes, conforme ilustrado na Figura 16. A primeira avaliação de desempenho foi realizada com base nos resultados obtidos por meio da técnica de validação cruzada, conforme descrito na seção anterior. Para essa análise, foram utilizadas métricas de avaliação do tipo macro, que consideram o desempenho médio entre as classes.

Figura 16 - Organização das avaliações de desempenho do modelo.



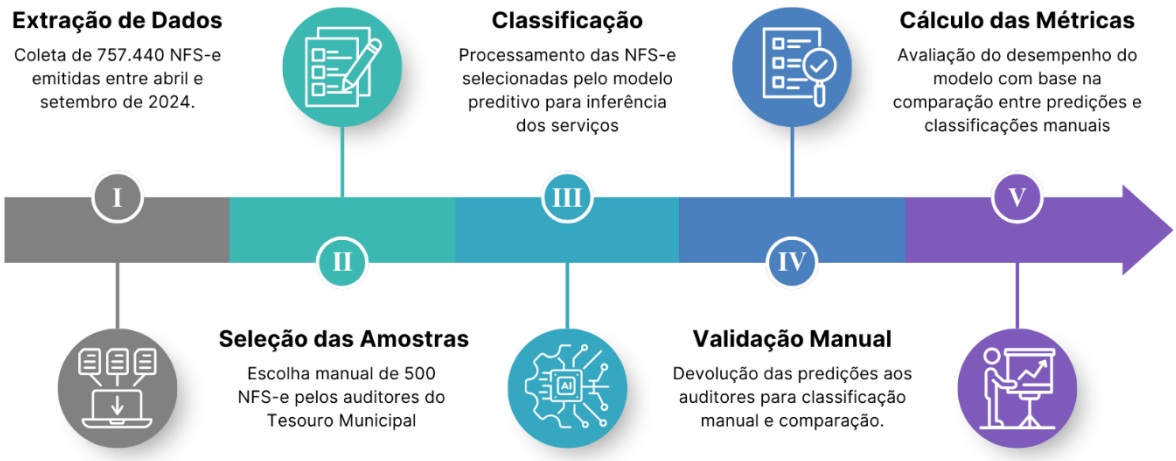
Fonte: Elaborado pelo autor

Já a segunda avaliação foi realizada exclusivamente com os quatro algoritmos que apresentaram os melhores desempenhos na etapa anterior. Nessa fase, os modelos foram treinados utilizando a totalidade da base de dados de treinamento, previamente balanceada, visando maximizar o aproveitamento das informações disponíveis. A avaliação contou, ainda, com a participação de três auditores do Tesouro Municipal da Prefeitura do Recife–PE.

Conforme apresentado na Figura 17, essa fase foi estruturada em cinco etapas. (I) Inicialmente, foram extraídas 757.440 NFS-e, emitidas entre os meses de abril e setembro de 2024. Esses dados foram exportados para planilhas eletrônicas e disponibilizados aos auditores.

(II) Em seguida, os auditores selecionaram um conjunto de 500 NFS-e, que foi encaminhado para processamento e classificação pelos modelos. (III) As notas selecionadas foram então submetidas ao modelo preditivo, que realizou a inferência dos serviços correspondentes. (IV) Posteriormente, as NFS-e foram devolvidas aos auditores contendo os serviços classificados automaticamente pelo modelo, possibilitando a inclusão e posterior comparação com as classificações manuais realizadas pelos especialistas. (V) Por fim, as métricas de avaliação foram calculadas a partir da comparação entre as predições geradas pelos modelos e as classificações manuais fornecidas pelos auditores.

Figura 17 - Organização dos testes em conjunto com Auditores.



Fonte: Elaborado pelo autor

Com base nos resultados obtidos nas duas etapas de avaliação, foi possível selecionar o algoritmo com melhor desempenho para compor a versão final do modelo preditivo do Nota Conforme. Em seguida, os resultados consolidados e o modelo selecionado foram apresentados à Secretaria de Finanças do Município do Recife–PE, com o objetivo de validar sua eficácia prática e discutir a viabilidade e os meios para sua implantação no contexto institucional. A temática relacionada à implementação do modelo será abordada em maior detalhe na seção seguinte.

4.7. Implantação do modelo

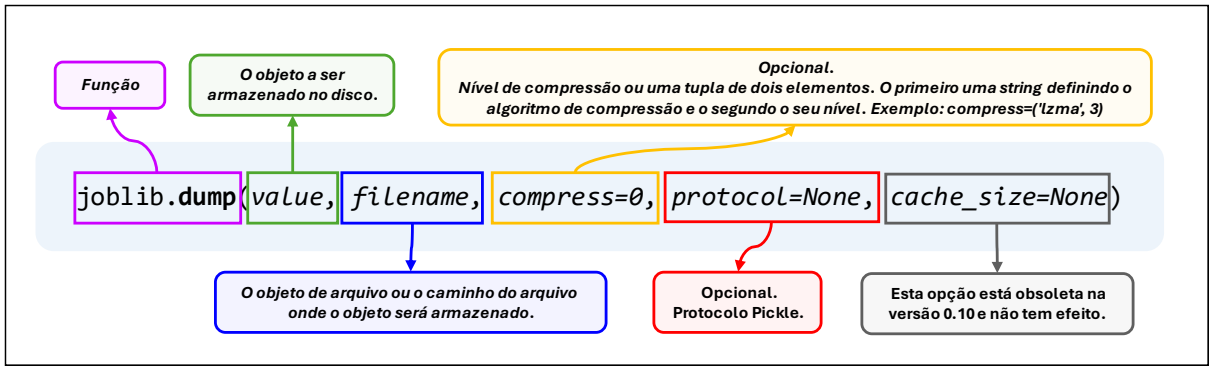
Com o modelo treinado e validado, a próxima etapa é a preparação da solução para viabilizar a sua implantação em ambiente de produção (GÉRON, 2021, p. 65). Dessa forma, essa sessão se concentra em detalhar esse processo de preparação do modelo e desenvolvimento da solução que viabilizará a disponibilização de novas predições em ambiente de produção.

4.7.1. Salvando o modelo treinado e a matriz TF-IDF em disco

Antes da implantação é preciso salvar o modelo treinado. Para isso, uma das alternativas é a utilização da biblioteca do *Python* chamada *Joblib* (GÉRON, 2021, p. 65). A *Joblib* é uma robusta biblioteca que utiliza P*ython* para implementar um conjunto de ferramentas para criação de *pipelines*, ela é projetada para operar com alta performance, mesmo em grandes volumes de dados (KALINOWSKI et al, 2023). A partir dessa biblioteca, é possível salvar o modelo treinado em disco e carregá-lo em diferentes ambientes, como, por exemplo, no ambiente de produção.

Portanto, o modelo escolhido e sua matriz vetorizada TF-IDF foram persistidos em disco utilizando a função *dump* da biblioteca. Essa função, segundo Joblib Developers (2024), possui cinco parâmetros, conforme apresentado na Figura 18. Os parâmetros *value* e *filename* são obrigatórios e definem, respectivamente, o objeto *python* a ser armazenado e o caminho do arquivo onde o objeto será armazenado.

Figura 18 - Detalhamento da função *dump*() da biblioteca *Joblib*.



Fonte: Adaptado de Joblib Developers (2024)

Como esta pesquisa utilizou um grande conjunto de dados para treinamento do modelo, optou-se pela otimização do uso de espaço em disco, aplicando o parâmetro *compress* para comprimir o modelo. Segundo a Joblib Developers (2024), esse parâmetro permite especificar tanto o algoritmo compressor (como, por exemplo, ZLIB, GZIP, BZ2, LZMA, XZ) quanto o seu nível de compressão, que varia de 0 a 9. O algoritmo de compressão LZMA foi o escolhido, utilizando o nível 3 de compressão, considerado um bom equilíbrio entre eficiência e taxa de compressão pela documentação oficial da biblioteca. Ao finalizar, foi gerado um modelo com tamanho de 2,65 g*igabytes* (GB) e uma matriz TF-IDF de 539 *kilobytes* (KB). Com isso, o modelo ficou pronto para implantação em ambiente de produção.

4.7.2. Desenvolvimento da *Application Programming Interface* (API)

A implantação do modelo preditivo em ambiente de produção é uma das últimas etapas do processo de desenvolvimento do sistema de software inteligente. Entre as formas de implantação, o modelo pode ser disponibilizado por meio de um serviço *web*, onde o modelo é carregado em um código *back-end* e as novas previsões são geradas conforme a demanda. Esse consumo pode ser feito, mas não se limita, a partir de uma aplicação *front-end* ou de integração com outros sistemas (KALINOWSKI et al., 2023).

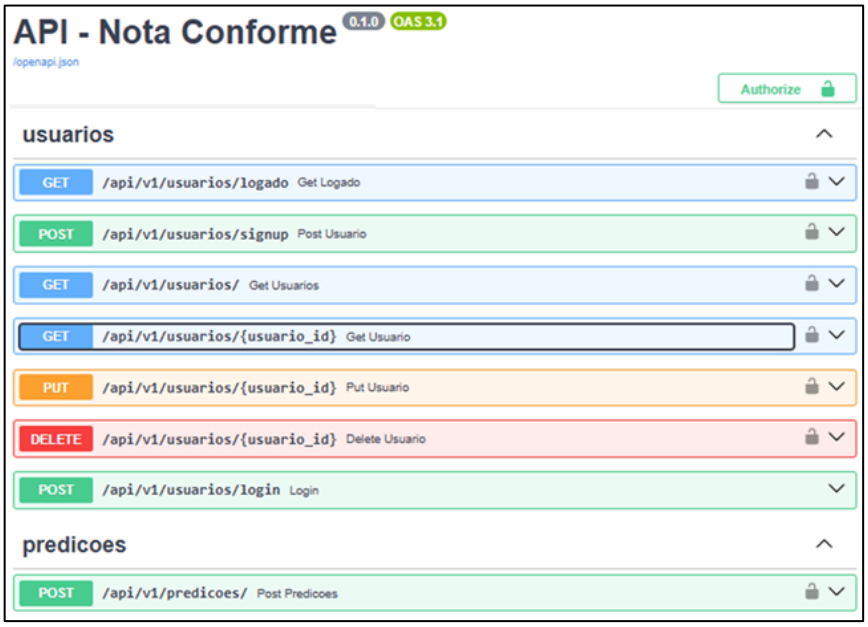
Uma das vantagens de utilizar esse formato de implantação é que ele permitirá que os aplicativos que consumirem os recursos da API não fiquem limitados à linguagem *Python*, permitindo aplicações desenvolvidas em qualquer linguagem (GÉRON, 2021, p. 66). Além disso, segundo Kalinowski et al. (2023), esse formato de implantação pode ser realizado na própria infraestrutura da organização, que oferece mais controle sobre o modelo e os dados utilizados, mas que depende de investimentos em hardware e infraestrutura para manter o serviço. Ou ainda pode ser hospedado em um provedor de serviços em nuvem, oferecendo maior escalabilidade e custo-benefício. No entanto, essa abordagem exige uma atenção especial aos aspectos de segurança.

Nesse sentido, para viabilizar a implantação do modelo em produção para automatização das predições de novas NFS-e, foi desenvolvida uma API utilizando o *framework* FastAPI. Segundo o seu criador, Ramírez (2024), o FastAPI é um *framework web* em Python para desenvolvimento de APIs modernas e de alto desempenho, com performance comparável à de soluções desenvolvidas em linguagens como Node.js e Go. Além disso, o framework passa por contínuas atualizações para aprimorar suas funcionalidades e corrigir falhas reportadas pela comunidade. Dispõe ainda de documentação atualizada em seu site oficial, dando mais confiabilidade para implantação em ambientes de produção.

Outra vantagem em utilizar o FastAPI é a geração automática da documentação da API contendo todas as suas rotas e seus respectivos *endpoints* utilizando a plataforma Swagger UI. Por padrão, o acesso à documentação pode ser realizado através do endereço http://localhost:8000/docs (RAMÍREZ, 2024). Ao acessá-la, além de visualizar toda a estrutura da API, conforme apresentado na Figura 19, é possível consultar os requisitos necessários para realizar requisições aos recursos e entender o formato das respostas que serão devolvidas.

Adicionalmente, o usuário pode realizar testes por meio de requisições reais, permitindo uma interação prática com a API.

Figura 19 - Documentação da API para predição das NFS-e no Swagger.



Fonte: Elaborado pelo autor

A camada de segurança da API foi implementada utilizando o *JSON Web Token* (JWT). Segundo o site oficial, JWT.io (2024), o JWT é um padrão aberto para transmissão segura de informações entre partes. O seu uso mais comum é para autenticação, como em sistemas de autenticação e autorização de API. Nesse cenário, quando o usuário faz *login*, o servidor gera um *token* JWT contendo informações do usuário e da autorização de acesso. Esse *token* é então retornado na resposta da requisição de autenticação e possui validade de 30 dias. Assim, o usuário poderá utilizá-lo em requisições futuras para autenticar-se e acessar os recursos aos quais o *token* concede permissão, até o seu vencimento. Após o vencimento do *token*, será necessário acessar novamente o recurso *login* para obtenção de um novo *token* válido.

4.8. Considerações finais

Este capítulo apresentou os métodos e técnicas utilizadas para desenvolver o Nota Conforme. Foram detalhadas as etapas de aquisição e pré-processamento dos dados, sua transformação em formato processável por computadores por meio de TF-IDF, seguida da construção, avaliação e escolha do modelo preditivo. Além disso, detalharam-se os aspectos técnicos do desenvolvimento da API como método de implantação e integração do modelo preditivo.

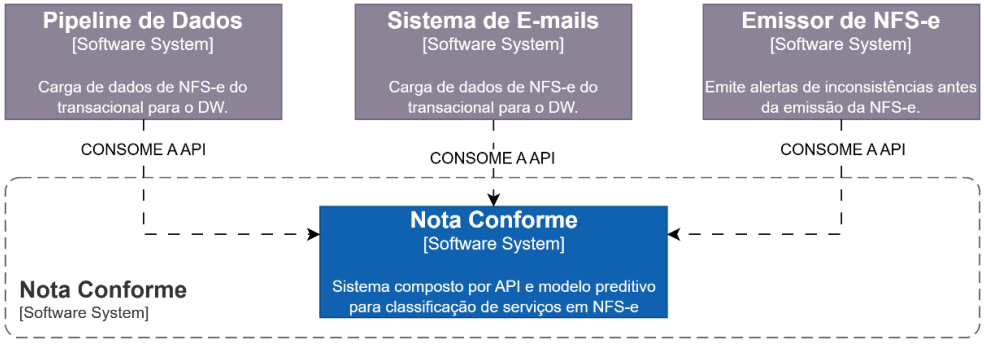
5. NOTA CONFORME: SISTEMA PARA CLASSIFICAÇÃO AUTOMATIZADA DE SERVIÇOS EM NFS-E COM MACHINE LEARNING

Este capítulo apresenta o Nota Conforme, sistema composto por um modelo preditivo e uma *Application Programming Interface* (API), desenvolvido para a Prefeitura do Recife–PE. Seu objetivo é integrar-se aos sistemas internos para automatizar a classificação de serviços em Notas Fiscais de Serviços Eletrônicas (NFS-e) e apoiar os auditores fiscais na identificação de inconsistências entre os serviços declarados e as alíquotas aplicadas durante as fiscalizações.

5.1. Arquitetura do Sistema

Visando atender as atuais necessidades e demandas futuras, a solução foi pensada para possibilitar a fácil integração com sistemas e *pipelines* de dados existentes na Prefeitura do Recife–PE e em seus órgãos controladores. De acordo com C4 Model (2024), o Diagrama de Contexto é responsável por apresentar uma visão geral do sistema e seu cenário de uso, exibindo os usuários e sistemas que interagem com a solução. O diagrama apresentado na Figura 20, ilustra a solução de forma geral, destacando as ferramentas e sistemas integrados ao Nota Conforme: o *pipeline* de dados, sistema de envio de e-mails e ao emissor de NFS-e.

Figura 20 - Diagrama de Contexto do Sistema Nota Conforme.

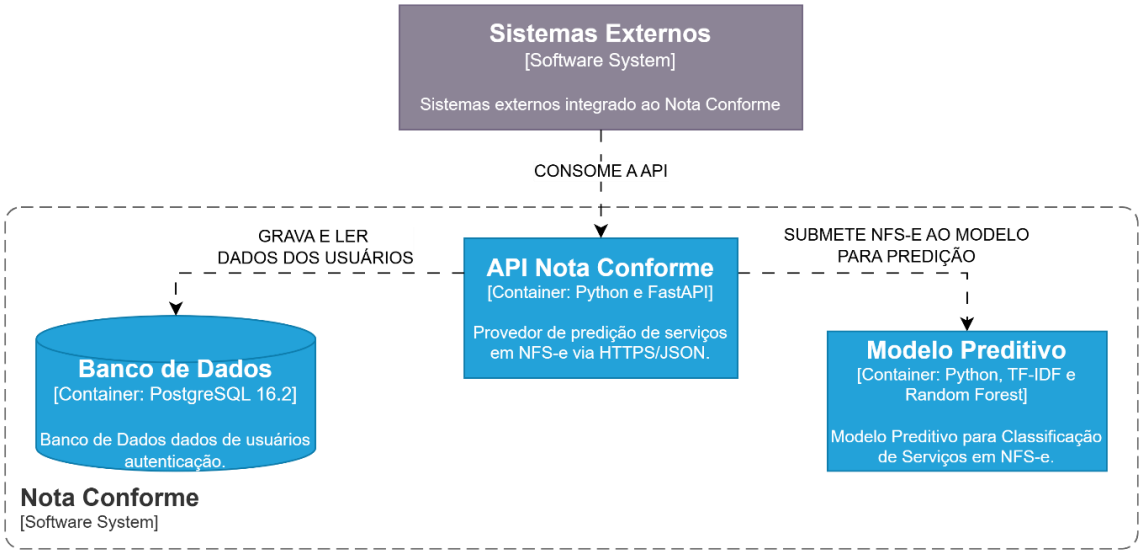


Fonte: Elaborado pelo autor

Já o Diagrama de Contêiner, visa ampliar a visão do sistema. Um contêiner pode ser considerado como uma unidade que executa códigos ou armazena dados separadamente. Esse tipo de diagrama ilustra a distribuição das responsabilidades, tecnologias escolhidas e como a comunicação é realizada entre os contêineres (C4 MODEL, 2024). A partir do Diagrama de Contêiner ilustrado na Figura 21, é possível identificar existência de três contêineres pertencentes ao Nota Conforme.

O primeiro contêiner, denominado de API, foi desenvolvido em Python utilizando o *framework* FastAPI. Ele é responsável pela integração com os sistemas externos, sendo o único que realiza esse tipo de comunicação. Ao receber uma requisição para realizar predições de serviços em NFS-e, a API processa os dados recebidos e submete-os ao contêiner Modelo Preditivo para o modelo realizar a identificação dos códigos de serviços. Quando se tratar de uma requisição inerente aos usuários, a API executa operações de consulta e manipulação de dados no banco de dados da aplicação.

Figura 21 - Diagrama de Contêiner do Sistema Nota Conforme.



Fonte: Elaborado pelo autor

O contêiner Banco de Dados, na versão atual do sistema, foi desenvolvido em PostgreSQL na versão 16.2. A base de dados é composta por uma única tabela, responsável pelo armazenamento de dados cadastrais e de autenticação dos usuários, conforme ilustrado pela Figura 22. O campo correspondente ao e-mail é utilizado como identificador de *login*, sendo combinado com a senha para autenticação no sistema. No contexto do sistema, a comunicação com o banco de dados é realizada somente através da API.

Figura 22 - Esquema banco de dados API.

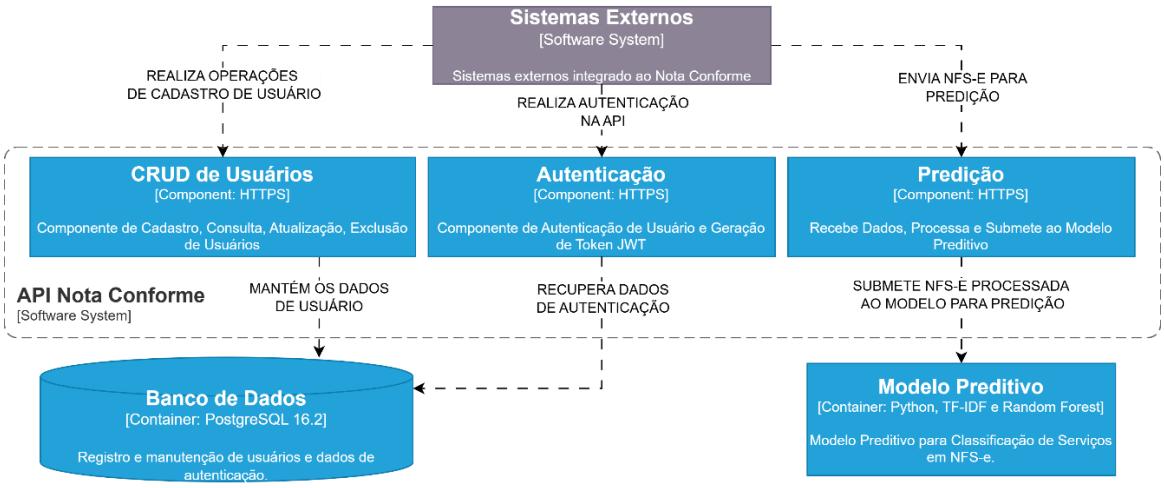
usuarios		
PK	id	bigint
	nome	varchar
	email	varchar
	senha	varchar
	ativo	bool
	created_at	timestamp
	updated_at	timestamp

Fonte: Elaborado pelo autor

A partir da construção do Diagrama de Componentes, é possível expandir a visão dos contêineres, detalhando os componentes que os integram, suas funções e responsabilidades no sistema, bem como as tecnologias empregadas em sua implementação (C4 MODEL, 2024). Ou seja, a partir desse tipo de diagrama é possível entender como o sistema está organizado por dentro. Conforme a Figura 23, por exemplo, conclui-se que o contêiner API é formado por três componentes:

- CRUD[1] de Usuários: Responsável pelas operações de criação, leitura, atualização e exclusão de dados de usuários.
- Autenticação: Executa as operações de autenticação do usuário, como validação de dados de *login* e geração de *token* do tipo *JSON Web Token* (JWT).
- Predição: Responsável pelo recebimento, processamento e submissão das NFS-e ao modelo preditivo para identificação dos serviços descritos nas NFS-e.

Figura 23 - Diagrama de Componente Nota Conforme.



Fonte: Elaborado pelo autor

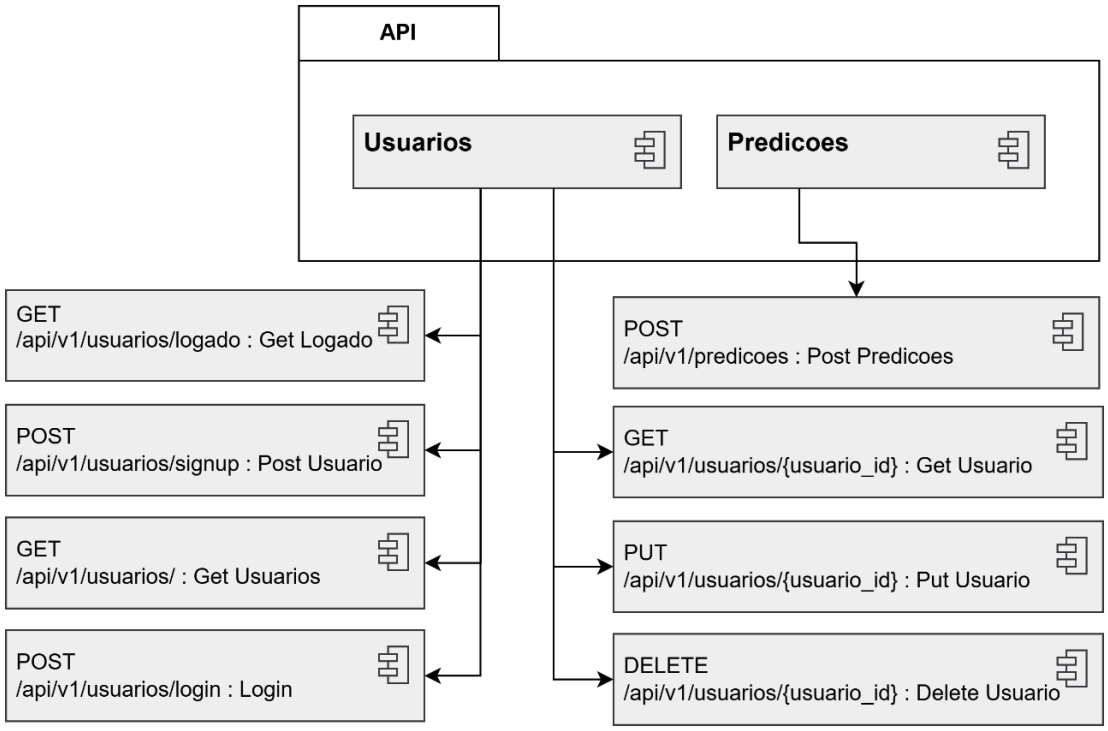
5.2. Funcionalidades do Sistema

Conforme ilustrado na Figura 24, o Nota Conforme apresenta dois recursos principais: **usuarios** e **predicoes**. O recurso de usuários está relacionado à gestão de dados dos usuários, contemplando funcionalidades como cadastro, atualização, exclusão, recuperação e autenticação (*login*). Já o recurso de predições, que constitui o núcleo da API, pretende processar um conjunto de NFS-e, realizando a classificação dos serviços prestados com base na

[1] *Create, Read, Update e Delete* (CRUD): É um acrônimo do inglês para as operações básicas de criação, leitura, atualização e exclusão de dados.

discriminação de cada nota. A figura também detalha os *endpoints* e os métodos *Hypertext Transfer Protocol* (HTTP) utilizados por cada recurso.

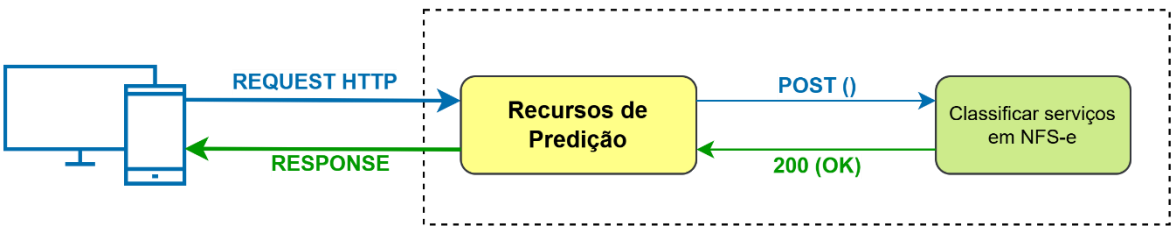
Figura 24 – Visão geral dos serviços, rotas e métodos da API Nota Conforme.



Fonte: Elaborado pelo autor

Já a Figura 25, além de apresentar o objetivo do recurso de predição da API, exibe graficamente o fluxo de requisição e resposta para predição de novas NFS-e. Além disso, a imagem reforça o método utilizado para requisição e indica o *status* da resposta de uma requisição processada com sucesso pela API.

Figura 25 – Status resposta de sucesso do recurso de predição da API Nota Conforme.

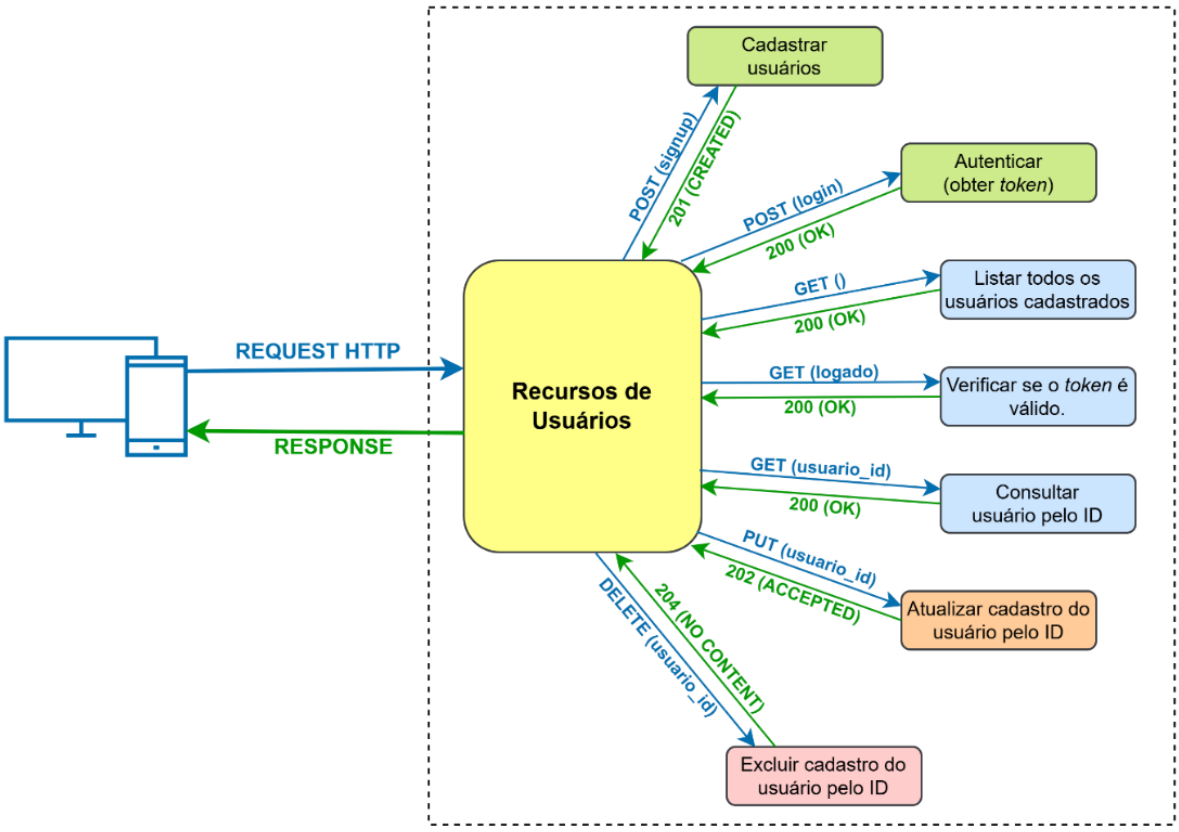


Fonte: Elaborado pelo autor

Com o mesmo objetivo da imagem anterior, a Figura 26 apresenta os objetivos, o fluxo de dados e o *status* da resposta de cada *endpoint* do recurso Usuários. Por exemplo, ao tentar

cadastrar um novo usuário por meio do método POST ao recurso *signup*, será devolvido o *status* de código 201 (*created*) em caso de sucesso.

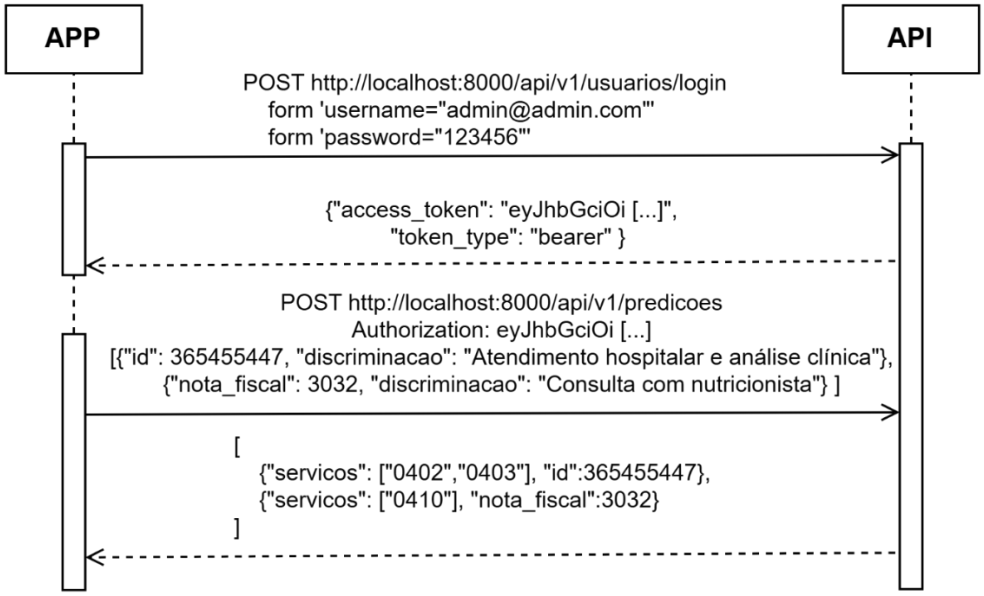
Figura 26 - Status resposta de sucesso do recurso de usuários da API Nota Conforme.



Fonte: Elaborado pelo autor

O diagrama de sequência, apresentado na Figura 27, ilustra de forma mais prática o funcionamento do Nota Conforme, destacando as etapas necessárias para o uso das rotas protegidas. Inicialmente, o usuário realiza a autenticação por meio de uma requisição **POST** à rota **/login**, enviando o nome de usuário e a senha no corpo da requisição, utilizando o formato ***form***. Como resposta, a API retorna um *token* JWT do tipo *Bearer*, com validade de 30 dias. Com o *token*, o usuário pode utilizá-lo como método de autenticação para acessar a rota **/predicoes**. Nessa requisição, também são enviados os dados das NFS-e a serem classificadas. Por fim, a API responde com as predições geradas pelo modelo, classificando os serviços com base nas informações fornecidas.

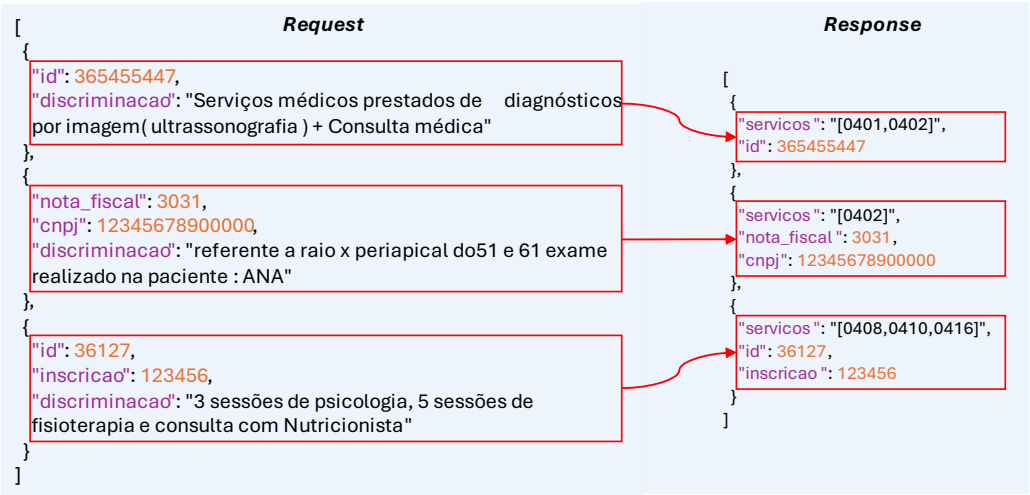
Figura 27 - Diagrama de sequência para requisição na API.



Fonte: Elaborado pelo autor

Para realizar predições de serviços em NFS-e, o usuário deverá incluir na requisição, no mínimo, o atributo *discriminacao*, que contém a discriminação da NFS-e. Para garantir a flexibilidade para possibilitar utilização do Nota Conforme em diferentes cenários, os demais campos são dinâmicos e opcionais, podendo ser informados os atributos que identificam unicamente a NFS-e em cada organização. Como resultado, o Nota Conforme responderá com o atributo *servicos*, que contém todos os códigos de serviços identificados na descrição da NFS-e pelo modelo preditivo, além dos demais campos enviados na requisição, conforme apresentado na Figura 28.

Figura 28 - Exemplo de requisições e respostas de predições por meio da API.



Fonte: Elaborado pelo autor

O único campo omitido na resposta é o atributo *discriminacao*, uma vez que ele não desempenha função na identificação da nota e sua exclusão ajuda a reduzir o volume de dados trafegados. Isso significa que, ao enviar uma requisição com os atributos *nota_fiscal*, *cnpj* e *discriminacao*, a API retornará os atributos *servicos*, *nota_fiscal* e *cnpj* no formato apresentado na Figura 28. Com isso, espera-se uma solução dinâmica capaz de facilitar a integração com diferentes sistemas.

5.3. Disponibilização do código-fonte e materiais do Nota Conforme

Os códigos e manuais desenvolvidos estão disponíveis para *download* em um repositório público do GitHub. Estes recursos podem ser utilizados livremente pela comunidade, e aprimorados ou adaptados conforme necessidade. Além disso, foi disponibilizado um vídeo demonstrativo do Nota Conforme. Esses materiais podem ser acessados através dos endereços disponíveis na Tabela 17.

Tabela 17 – Código-fonte e Vídeo Demonstrativo do Nota Conforme.

Material	Endereço
Código-fonte	https://github.com/tpfarias/notaconforme
Vídeo Demonstrativo	https://youtu.be/OLaRtmaSyXY

Fonte: Elaborado pelo autor

O sistema Nota Conforme foi submetido à sessão Demos e Aplicações do 40º Simpósio Brasileiro de Banco de Dados (SBBD) por meio do artigo apresentado no APÊNDICE B, acompanhado de seu respectivo código-fonte e vídeo demonstrativo. A submissão foi aceita, conforme evidenciado pelo e-mail disponível no ANEXO A.

5.4. Considerações finais

Este capítulo apresentou a solução Nota Conforme. Foi discutida sua estrutura arquitetural por meio dos diagramas de contexto, contêineres e componentes. Além disso, foram demonstradas as suas funcionalidades, bem como as etapas necessárias para consumir as rotas da API do Nota Conforme, incluindo um exemplo de requisição e a resposta esperada. O próximo capítulo visa analisar os resultados do Nota Conforme, a partir de uma discussão detalhada sobre o desempenho do modelo preditivo e funcionamento da API.

6. RESULTADOS E DISCUSSÕES

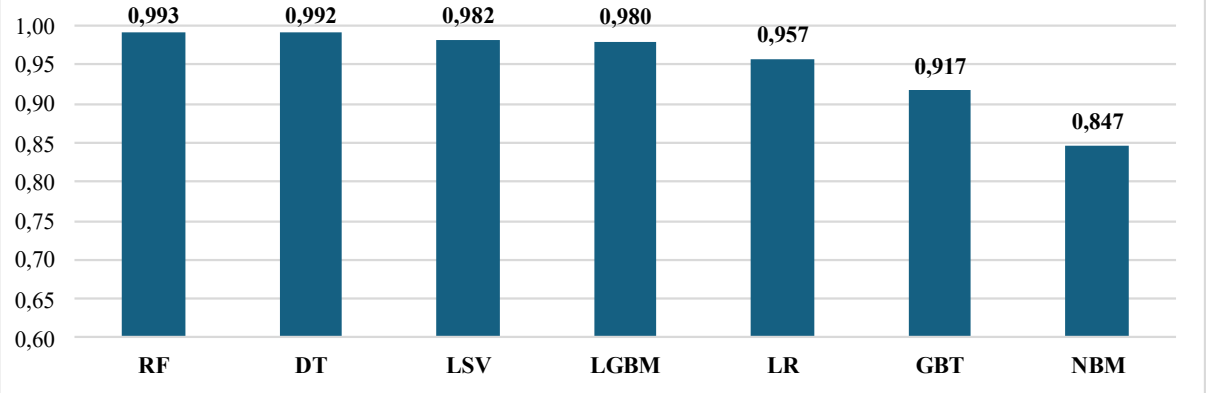
Este capítulo tem como objetivo apresentar e analisar os resultados obtidos ao longo desta pesquisa. Na primeira parte, discute-se o desempenho dos modelos desenvolvidos, que serviram de base para a escolha do modelo adotado no Nota Conforme. Em seguida, é detalhada a *Application Programming Interface* (API) de integração criada para viabilizar a implantação do modelo em ambiente de produção. Por fim, são discutidos os resultados da avaliação de viabilidade a aceitação do Nota Conforme, a partir das percepções dos potenciais usuários.

6.1. Avaliação comparativa dos modelos de classificação

Esta seção descreve e discute os resultados obtidos na avaliação comparativa dos algoritmos aplicados à tarefa de classificação multirrótulo. Foram analisados os algoritmos: *Random Forest* (RF), *Decision Tree* (DT), *Linear Support Vector* (LSV), *LightGBM* (LGBM), *Logistic Regression* (LR), *Gradient Boosting* (GBT) e *Naive Bayes Multinomial* (NBM).

A Figura 29 evidencia que o modelo *Random Forest* obteve a maior acurácia entre os avaliados, alcançando 99,3%. O *Decision Tree* veio em seguida com 99,2%. Já o *Linear Support Vector* e o *LightGBM* alcançaram 98,2% e 98,0%, respectivamente. Contudo, é importante destacar que, dada a natureza multirrótulo do problema, a acurácia isoladamente não constitui um indicador suficiente para uma avaliação conclusiva. Dessa forma, outras métricas foram consideradas para uma análise mais robusta.

Figura 29 - Análise Comparativa da Acurácia dos Modelos Treinados



Fonte: Elaborado pelo Autor

A avaliação das demais métricas, apresentadas na Tabela 18, confirmou o desempenho superior do modelo *Random Forest*, com valores elevados de precisão (0,9978), *recall* (0,9960) e *F1-score* (0,9969), além do menor índice de *Hamming Loss* (HL) de 0,00039. Tais resultados

indicam boa capacidade preditiva e uma forte robustez na classificação multirrótulo. No entanto, esses benefícios vêm acompanhados de custos computacionais expressivos: o modelo apresentou o maior tempo médio de treinamento (TMT), de 01:47:36, e o maior tempo médio de predição (TMP), de 00:01:29. Esse fator pode limitar sua aplicabilidade em cenários que demandem alta eficiência computacional ou respostas em tempo real.

O modelo *Decision Tree*, embora também apresente um TMT elevado (01:45:05), destacou-se pelo baixo TMP (00:00:07). Suas métricas de desempenho também foram elevadas, incluindo um *F1-score* de 0,9965 e um HL de 0,00045, configurando-se como uma alternativa competitiva em termos de precisão, com menor custo na fase de inferência.

Tabela 18 – Análise comparativa dos modelos

Modelo	Precisão	Recall	F1-Score	HL	TMT	TMP
Random Forest	0,9978	0,9960	0,9969	0,00039	01:47:36	00:01:29
Decision Tree	0,9971	0,9958	0,9965	0,00045	01:45:05	00:00:07
Linear Support Vector	0,9948	0,9914	0,9931	0,00093	00:10:28	00:00:03
LightGBM	0,996	0,9901	0,9930	0,00101	00:15:06	00:00:12
Logistic Regression	0,9893	0,9758	0,9825	0,00220	00:01:09	00:00:03
Gradient Boosting	0,9917	0,9553	0,9719	0,00400	03:46:38	00:00:10
Naive Bayes Multinomial	0,9168	0,9490	0,9296	0,00921	00:00:05	00:00:04

Fonte: Elaborado pelo Autor

Por sua vez, o modelo *Linear Support Vector* demonstrou um equilíbrio notável entre desempenho e eficiência computacional. Com *F1-score* de 0,9931 e HL de 0,00093, foi o modelo mais rápido tanto no treinamento (00:10:28) quanto na predição (00:00:03) entre os quatros mais bem avaliados.

De forma semelhante, o *LightGBM* apresentou bom desempenho preditivo, com *F1-score* de 0,9930 e HL de 0,00101, destacando-se também pela eficiência computacional, com TMT de 00:15:06 e TMP de 00:00:12. Essa combinação de assertividade e baixos custos computacionais torna ambos os modelos – *Linear Support Vector* e *LightGBM* – particularmente adequados para cenários que exigem um equilíbrio entre qualidade preditiva e limitações de tempo ou recursos computacionais.

O modelo *Logistic Regression*, apesar de apresentar métricas inferiores em relação aos modelos previamente discutidos, ainda demonstrou resultados satisfatórios. Com *F1-score* de 0,9825 e HL de 0,00220, o modelo manteve um bom desempenho para a tarefa proposta. Seu principal diferencial reside na baixa complexidade computacional: com TMT de somente

00:01:09 e TMP de 00:00:03, trata-se de uma solução extremamente leve, podendo ser adequada para cenários no qual os recursos computacionais são limitados, mesmo que isso implique em uma possível perda de qualidade preditiva.

Em contraste, o modelo *Naive Bayes Multinomial*, apesar de extremamente eficiente em termos de tempo de treinamento (00:00:05), apresentou limitações significativas de desempenho, com um *F1-score* de 0,9296, o mais baixo entre os modelos avaliados. Tal resultado sugere que esse algoritmo pode não ser adequado para o contexto em questão.

Por fim, o modelo *Gradient Boosting* alcançou um *F1-score* de 0,9719, considerado satisfatório. No entanto, seu tempo médio de treinamento, de 03:46:38, foi o mais elevado dentre todos os modelos, comprometendo sua viabilidade em aplicações com restrições de tempo ou recursos computacionais.

Esses resultados evidenciam a importância de considerar não somente a precisão preditiva, mas também os aspectos relacionados à eficiência computacional, especialmente em cenários práticos que demandem escalabilidade e respostas em tempo hábil. Diante disso, os quatro algoritmos mais bem avaliados (*Random Forest, Decision Tree, Linear Support Vector* e *LightGBM*) foram escolhidos para a próxima etapa de testes, considerando seu desempenho equilibrado entre a qualidade preditiva e viabilidade computacional. Por fim, os resultados e análises apresentados nesta seção serviram de base para a elaboração do artigo incluído no APÊNDICE A.

6.2. Avaliação a partir da base selecionada pelos auditores

No segundo experimento, os quatro algoritmos selecionados na etapa anterior foram reavaliados utilizando uma amostra de 500 Nota Fiscal de Serviço Eletrônica (NFS-e), escolhidas por auditores da Prefeitura do Recife–PE. O objetivo foi avaliar o desempenho real dos modelos, utilizando um conjunto de dados que refletisse com mais fidelidade os desafios enfrentados pelos auditores em suas análises rotineiras de fiscalização, como a identificação de irregularidades sutis e padrões suspeitos em meio a grandes volumes de documentos fiscais.

A *Random Forest* continuou sendo o modelo com melhor performance, conforme ilustrado na Tabela 19. Com uma precisão de 0,941, *recall* de 0,952 *e F1-Score* de 0,945, o modelo demonstrou maior robustez à variação dos dados. O alto *F1-Score* indica um bom equilíbrio entre precisão e *recall*, o que o torna bastante confiável para a tarefa. O *Linear*

Support Vector também teve um desempenho sólido, com um *F1-Score* de 0,924. Esse modelo se destacou por sua precisão (0,920) e *recall* (0,931) equilibrados, embora ligeiramente inferiores aos da *Random Forest*.

Tabela 19 – Desempenho dos modelos na base selecionado por auditores

Modelo	Precisão	Recall	F1-Score
Random Forest	0,941	0,952	0,945
Linear Support Vector	0,920	0,931	0,924
Decision Tree	0,911	0,924	0,915
LightGBM	0,921	0,905	0,903

Fonte: Elaborado pelo Autor

O *Decision Tree*, embora mais simples em estrutura, teve um desempenho competitivo, com um *F1-Score* de 0,915. Seus resultados (precisão de 0,911 e *recall* de 0,924) indicam que ele ainda é uma opção válida, especialmente em cenários onde a facilidade de interpretação e a rapidez na predição são fatores importantes, vantagens que se destacam quando comparado à maior complexidade do *Random Forest*. Por fim, o *LightGBM*, que havia tido bom desempenho inicial, apresentou o menor *F1-Score* (0,903) entre os quatro modelos. Seu *recall* (0,905) foi o mais baixo do grupo, o que pode indicar maior dificuldade em identificar corretamente os serviços descritos na NFS-e.

Esses resultados demonstram que, embora todos os modelos apresentem bom desempenho em ambientes controlados, o teste com dados reais evidencia diferenças mais claras em sua robustez e adequação prática. A *Random Forest*, em particular, se destaca por manter um desempenho elevado, mesmo diante da complexidade e imprevisibilidade dos dados das NFS-e. Apesar de apresentar o maior tempo de predição entre os modelos avaliados, esse tempo se mostra compatível com o ambiente de produção em que será implantado. Por esse motivo, o modelo foi escolhido para compor a atual versão do Nota Conforme, por oferecer o melhor equilíbrio entre desempenho, estabilidade e confiabilidade em cenários mais próximos da realidade enfrentada pelos auditores.

6.2.1. Avaliação dos erros de classificação do Randon Forest

Embora a avaliação quantitativa tenha demonstrado desempenho satisfatório do modelo Random Forest, com precisão de 94,1%, é importante refletir sobre os 5,9% de erros, que correspondem a 30 casos da amostra de 500 notas fiscais. A análise desses erros possibilita

compreender melhor os limites do modelo e identificar padrões de falha que não são captados pelas métricas tradicionais.

Ao inspecionar os erros manualmente, observou-se, a partir da Tabela 20, que 70% dos erros identificados ocorreram nas classes 0401 (20%), 0403 (26,67%) e 1 (23,33%). Essas classes apresentam grande variabilidade nos termos, pois as classes 0401 e 0403 abrangem uma ampla heterogeneidade de serviços. Já a classe 1, por se tratar de descrições genéricas que não estão vinculadas a um serviço específico, concentra muitas ocorrências de termos inéditos para o modelo, que podem levar a interpretações equivocadas sobre o serviço em questão.

Tabela 20 – Distribuição da frequência e proporção de erros por classe.

Código da Classe	Frequência de Erros	Percentual de Erros
0401	6	20,00%
0402	2	6,67%
0403	8	26,67%
0409	1	3,33%
0410	2	6,67%
0416	1	3,33%
0420	3	10,00%
1	7	23,33%

Fonte: Elaborado pelo Autor

Além disso, verificou-se a presença de termos característicos de outras classes em NFS-e que deveriam ser classificadas como genéricas, o que resultou em classificações equivocadas pelo modelo. A Tabela 21 ilustra alguns desses casos. Nos dois primeiros exemplos, o termo **médicos** aparece com frequência, na base de treinamento, associado às classes 0401, 0402 e 0403. Já no último exemplo, o termo **radiologia** ocorre predominantemente na classe 0402. Esses exemplos evidenciam que o modelo tende a apresentar dificuldades diante de termos ambíguos.

Tabela 21 – Exemplos de NFS-e da classe 1 com classificação incorreta.

Discriminação da NFS-e	Classe Correta	Classe Predita
REFERENTE AOS SERVIÇOS **MÉDICOS** PRESTADOS PELA DRA. [SUPRIMIDO], NO MES DE MAIO DE 2024.DADOS BANCARIOS PARA PAGAMENTO: BANCO: [SUPRIMIDO] AGENCIA: [SUPRIMIDO] CONTA: [SUPRIMIDO] CHAVE PIX: [SUPRIMIDO]	1	0403
REFERENTE AOS SERVIÇOS **MÉDICOS** PRESTADOS NA [SUPRIMIDO] NO MES DE JUNHO DE 2024, DRA. [SUPRIMIDO], NOS DIAS (03-06-10-13-17-20-24-27 - DIURNO). DADOS BANCÁRIOS PARA PAGAMENTO BANCO: [SUPRIMIDO] CÓD: [SUPRIMIDO] AGÊNCIA: [SUPRIMIDO] C/CORRENTE: [SUPRIMIDO] CHAVE PIX: [SUPRIMIDO]	1	0403

COORDENADOR DE **RADIOLOGIA** REFERENTE A MAIO/2024.VALOR APROXIMADO DE TRIBUTOS [SUPRIMIDO] FONTE: IBPT SEGUE DADOS BANCÁRIOS PARA TRANSFERÊNCIA OU DEPÓSITO: BANCO: [SUPRIMIDO] AGÊNCIA: [SUPRIMIDO] CONTA CORRENTE: [SUPRIMIDO] RAZÃO SOCIAL: [SUPRIMIDO] CNPJ: [SUPRIMIDO]	1	0402

Fonte: Elaborado pelo Autor

Outro aspecto identificado foi a presença do nome fantasia do prestador de serviço na discriminação da NFS-e. No exemplo ilustrado na Tabela 22, havia apenas a prestação de um serviço de fonoaudiologia, que deveria ser classificado unicamente como 0408. No entanto, como o nome fantasia continha os termos fonoaudiologia e neuropsicologia, o modelo equivocou-se ao atribuir também a classe 0416, correspondente a serviços da área de psicologia.

Tabela 22 – Exemplo de NFS-e com nome fantasia do prestador na discriminação.

Discriminação da NFS-e	Classe Correta	Classe Predita
Nome da Profissional: [SUPRIMIDO] Especialidade: Fonoaudiologia CRFª [SUPRIMIDO] Nome do(a) paciente: [SUPRIMIDO] Serviços prestados em FONOAUDIOLOGIA nos dias 5,7,12,14,19,21,26,28 de MARÇO de 2024. Valor por sessão: [SUPRIMIDO] DADOS BANCÁRIOS Banco [SUPRIMIDO] AGENCIA: [SUPRIMIDO] CONTA CORRENTE: [SUPRIMIDO] Favorecido - [SUPRIMIDO] **Fonoaudiologia** e **Neuropsicologia** [SUPRIMIDO]. CNPJ [SUPRIMIDO] PIX CNPJ [SUPRIMIDO]	0408	0408 e 0416

Fonte: Elaborado pelo Autor

Portanto, os erros observados demonstram que, apesar do bom desempenho geral, o modelo ainda apresenta limitações no tratamento de ambiguidades e de nuances linguísticas próprias do domínio fiscal. Essas restrições podem ser superadas em trabalhos futuros por meio da inclusão de técnicas adicionais na etapa de pré-processamento da base de treinamento, bem como pela adoção de recursos mais avançados, como *embeddings*, capazes de lidar de forma mais eficaz com essas particularidades.

A inspeção manual dos erros mostrou-se útil para compreender o comportamento do modelo nesta etapa da pesquisa. Entretanto, essa análise evidencia também a relevância de incorporar métodos específicos de explicabilidade, como LIME ou SHAP, a fim de fornecer interpretações mais sistemáticas e robustas sobre as decisões do modelo.

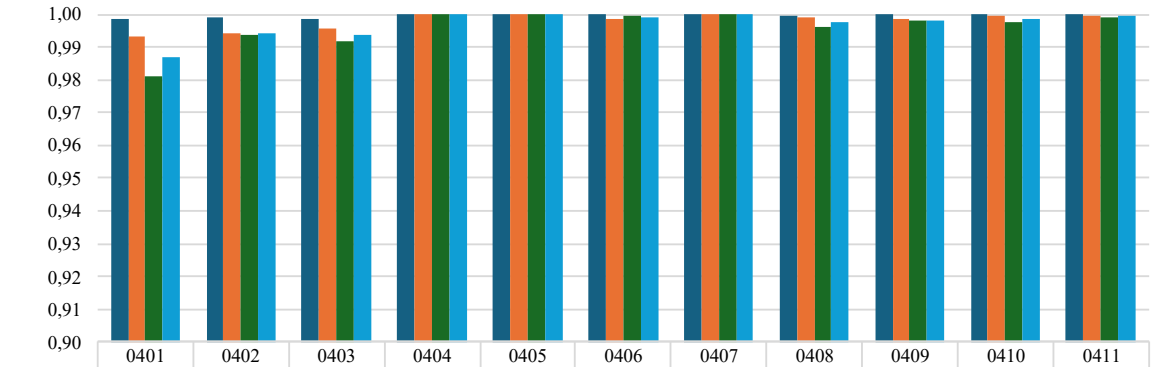
6.3. Avaliação de desempenho por classe do *Random Forest*

A partir da análise das métricas realizadas anteriormente, observou-se a robustez e a confiabilidade do modelo a partir de diferentes cenários de avaliação. Entretanto, é fundamental

analisar o desempenho individual de cada classe do modelo, uma vez que diferentes classes podem apresentar padrões e níveis de dificuldade distintos para o modelo, o que pode impactar diretamente sua aplicabilidade no contexto das NFS-e.

Os resultados apresentados nas figuras 30 e 31 mostram que o modelo apresenta um bom desempenho na classificação da maioria dos rótulos, evidenciado pelos altos valores de precisão, *recall*, *F1-score* e acurácia. Das 22 classes do modelo, 20 (90,9%) apresentaram *F1-score* superior a 0,99. Esse desempenho indica uma boa capacidade do modelo em identificar corretamente os casos positivos em praticamente todas as classes.

Figura 30 - Métrica de avaliação de desempenho do modelo por classe - Parte 1.



	0401	0402	0403	0404	0405	0406	0407	0408	0409	0410	0411
Acurácia	0,99849	0,99903	0,99857	0,99999	0,99999	0,99993	0,99999	0,99956	0,99980	0,99986	0,99994
Precisão	0,99303	0,99437	0,99558	0,99982	0,99997	0,99864	0,99996	0,99896	0,99867	0,99952	0,99961
Recall	0,98120	0,99366	0,99182	1,00000	0,99984	0,99974	0,99989	0,99593	0,99790	0,99772	0,99922
F1-Score	0,98708	0,99401	0,99370	0,99991	0,99991	0,99919	0,99993	0,99744	0,99828	0,99862	0,99942

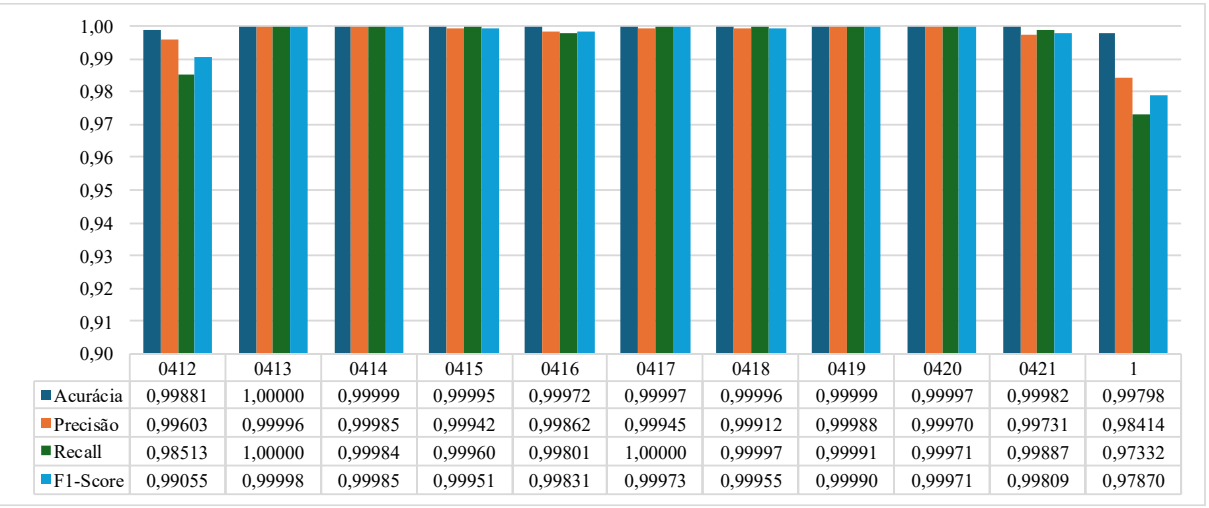
Fonte: Elaborado pelo autor

Por outro lado, embora o desempenho geral do modelo permaneça elevado, algumas classes apresentaram resultados ligeiramente inferiores. A classe 0401, por exemplo, obteve um *recall* de 0,098, o que impactou negativamente seu *F1-score*, resultando em 0,98. Esse contraste é relevante, sobretudo considerando que as demais classes atingiram métricas superiores a 0,99. O valor de *recall* indica uma dificuldade específica na identificação de exemplos positivos dessa classe, ou seja, o modelo apresentou um volume maior de falsos negativos (FN).

Outro destaque é a classe 1, que apresenta um *recall* de 0,97, indicando ainda uma boa performance, mas inferior em relação as demais classes. Apesar de possuir muitos registros para treinamento, por se tratar de uma descrição considerada genérica, as notas da classe 1 podem sofrer com a grande variabilidade de termos nas descrições. Com isso, alguns termos podem aparecer em quantidade insuficiente para que o modelo seja capaz de aprender a identificá-las.

Esses casos podem ser melhorados através da geração de novos dados, ajustes no pré-processamento ou técnicas de balanceamento para melhorar a representação dessas classes em versões futuras do modelo.

Figura 31 - Métrica de avaliação de desempenho do modelo por classe - Parte 2.

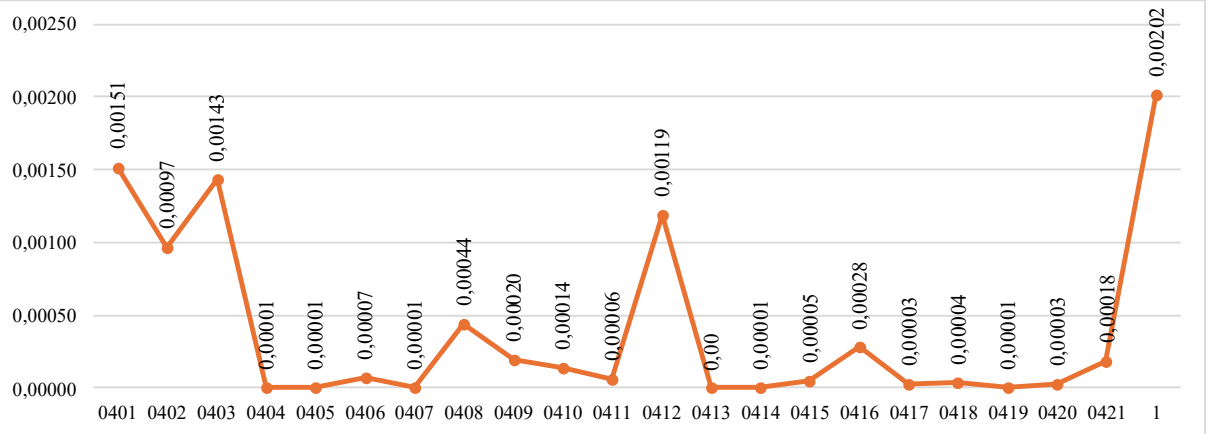


	0412	0413	0414	0415	0416	0417	0418	0419	0420	0421	1
Acurácia	0,99881	1,00000	0,99999	0,99995	0,99972	0,99997	0,99996	0,99999	0,99997	0,99982	0,99798
Precisão	0,99603	0,99996	0,99985	0,99942	0,99862	0,99945	0,99912	0,99988	0,99970	0,99731	0,98414
Recall	0,98513	1,00000	0,99984	0,99960	0,99801	1,00000	0,99997	0,99991	0,99971	0,99887	0,97332
F1-Score	0,99055	0,99998	0,99985	0,99951	0,99831	0,99973	0,99955	0,99990	0,99971	0,99809	0,97870

Fonte: Elaborado pelo autor

Além disso, a análise do *Hamming Loss* por classe reforça a variação de desempenho entre as categorias. A classe 0401 apresentou um dos maiores valores de *Hamming Loss*, atingindo 0,00151, indicando uma taxa de erro relevante na atribuição de rótulos e sugere que essa classe merece atenção especial. Da mesma forma, as classes, 0403 e 0412 apresentaram certo grau de instabilidade, com Hamming Loss de 0,00143 e 0,00119 respectivamente.

Figura 32 – Resultado de *Hamming Loss* por Classe.



Fonte: Elaborado pelo autor

A classe 1, com 0,00202, teve o maior valor entre todas, sugerindo a necessidade de atenção especial na modelagem dessa categoria. Em contraste, diversas classes, como 0404, 0405, 0407, 0414 e 0419, apresentaram *Hamming Loss* praticamente nulo (0,00001), refletindo um desempenho consistente e altamente preciso na classificação. Esses dados evidenciam que, embora o modelo tenha alcançado desempenho satisfatório em muitas classes, ainda há espaço para ajustes finos, especialmente nas classes com maior propensão a erros de classificação.

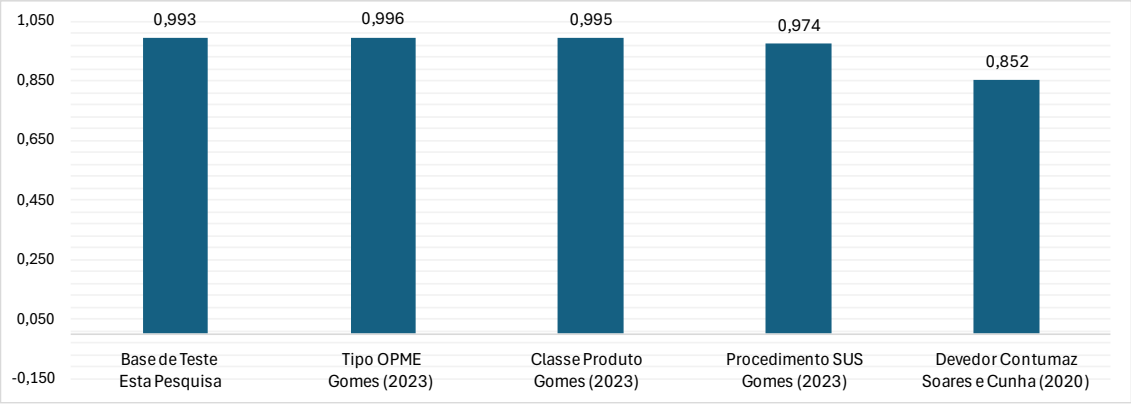
6.4. Desempenho em relação aos trabalhos relacionados

Esta seção apresenta os resultados de desempenho do modelo proposto, comparando-os com aqueles reportados em estudos anteriores. O objetivo é verificar se o desempenho do modelo desenvolvido está compatível com os resultados encontrados na literatura.

A Figura 33 apresenta a acurácia obtida pelo modelo deste estudo em relação a modelos desenvolvidos por Gomes (2023), que também utilizaram o algoritmo *Random Forest*. Observa-se que o desempenho do modelo atual foi semelhante aos modelos identificados como Tipo OPME e Classe Produto, sugerindo que a abordagem adotada nesta pesquisa está alinhada com boas práticas de aplicação do *Random Forest* em tarefas similares.

Já em comparação aos modelos Procedimento SUS, também de Gomes (2023), e Devedor Contumaz, de Soares e Cunha (2020), o modelo proposto apresentou uma acurácia superior. Contudo, é importante considerar que as diferenças podem estar relacionadas a fatores como o tipo e a qualidade dos dados utilizados, bem como às etapas de pré-processamento adotadas em cada estudo.

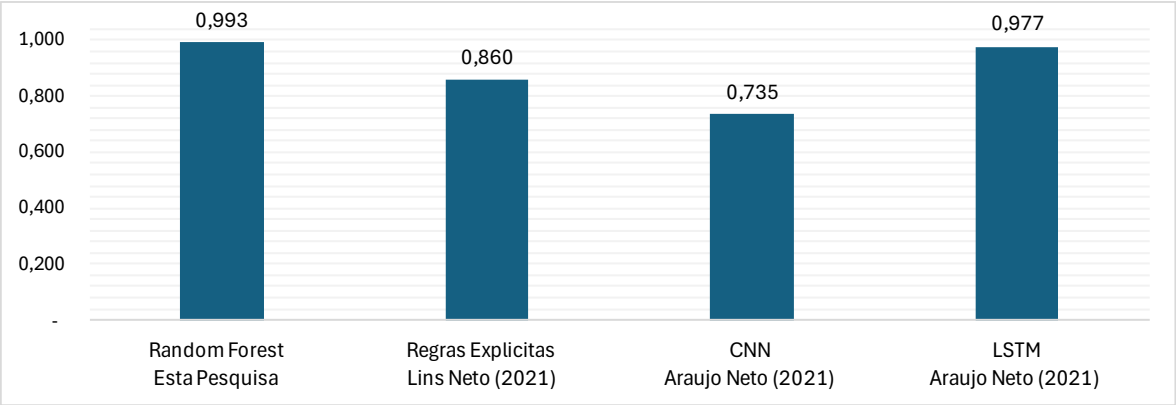
Figura 33 - Comparação da acurácia com trabalhos relacionados que utilizaram Random Forest.



Fonte: Elaborado pelo autor

A Figura 34 amplia a análise comparativa com os trabalhos similares, incluindo estudos que aplicaram outras técnicas ou algoritmos de aprendizado de máquina. Nessa comparação, o modelo desenvolvido nesta pesquisa obteve acurácia superior à de abordagens baseadas em regras explícitas e também superou modelos baseados em redes neurais, conforme reportado nos trabalhos analisados. Ainda que esses resultados indiquem um desempenho competitivo do modelo, ressalta-se que as comparações devem ser interpretadas com cautela, uma vez que os contextos de aplicação e as bases de dados diferem entre os estudos.

Figura 34 - Comparação da acurácia com trabalhos relacionados que utilizaram outras técnicas ou algoritmos.



Fonte: Elaborado pelo autor

Os resultados obtidos reforçam a viabilidade e o potencial do algoritmo *Random Forest* para a tarefa em questão. Além disso, observou-se que o desempenho do modelo desenvolvido é compatível com os resultados reportados na literatura, o que fortalece sua credibilidade. Dessa forma, conclui-se que o modelo apresenta desempenho satisfatório, demonstrando-se apto para aplicação em ambientes de produção.

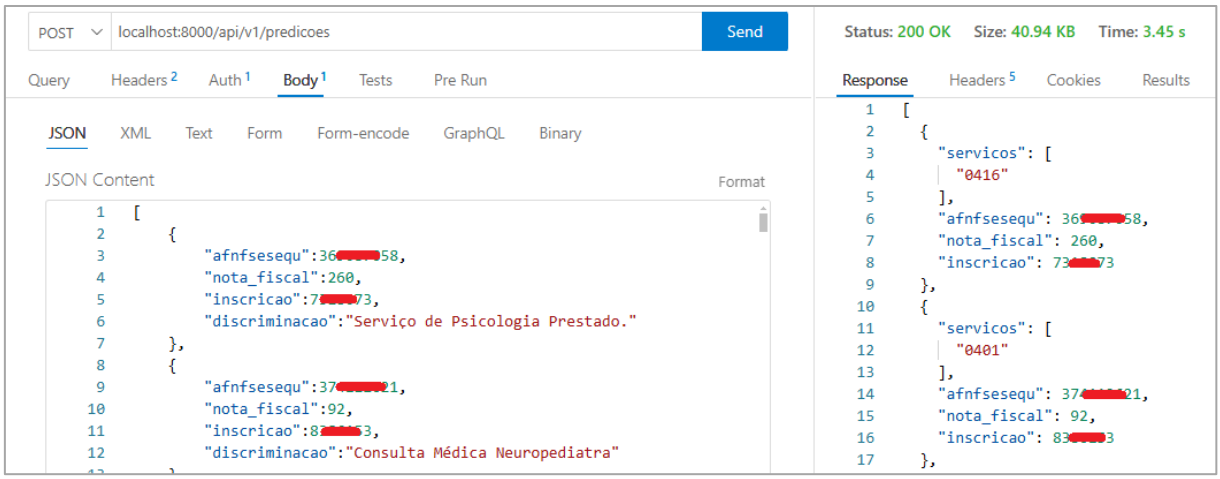
6.5. API de classificação de serviços em NFS-e

Além do desenvolvimento de um modelo para classificar serviços descritos em NFS-e, este estudo busca integrá-lo a um sistema que possibilite a automatização das predições e integração com outros sistemas. Como resposta, foi desenvolvido uma API utilizando o *framework* FastAPI do Python, conforme visto em capítulos anteriores.

Para avaliar o desempenho da API do Nota Conforme, foi realizada a submissão simultânea de 500 NFS-e para a API. O tempo total de processamento foi de 3.45 segundos, representando uma latência média de 6,9 milissegundos por NFS-e processada, conforme

apresentado na Figura 35. No âmbito da Prefeitura do Recife–PE, por exemplo, que registra uma média diária de 82 mil emissões de NFS-e, seriam necessários aproximadamente 9 minutos diariamente para processamento de todas as NFS-e emitidas no dia. Portanto, o sistema demonstra sua viabilidade para implantação em grandes conjuntos de dados.

Figura 35 - Resultado da requisição à API Nota Conforme.



Fonte: Elaborado pelo autor

Dessa forma, o Nota Conforme permite não apenas que a Prefeitura do Recife–PE automatize seu processo de classificação através da integração com sistemas internos, mas possibilita que qualquer prefeitura ou órgão controlador possa treinar seus próprios modelos e encapsulá-los ao Nota Conforme para automatização das predições. Isso é possível, pois esse formato de implantação viabiliza a integração a partir de soluções desenvolvida em qualquer linguagem de programação (GÉRON, 2021, p. 66).

6.6. Avaliação de viabilidade e aceitação do Nota Conforme

Esta seção apresenta a análise dos resultados obtidos por meio do questionário aplicado a auditores fiscais e a um gestor da administração tributária da Prefeitura do Recife, para avaliar a aceitação do sistema Nota Conforme. O instrumento de coleta de dados foi estruturado em duas etapas. A primeira etapa da pesquisa foi a caracterização do perfil dos participantes, com o propósito de avaliar suas experiências na análise de NFS-e, o nível de familiaridade com tecnologias aplicadas para fiscalização e suas percepções quanto às inovações tecnológicas incorporadas às rotinas de auditoria.

Já na segunda etapa do questionário, foram aplicadas questões com o propósito de avaliar a viabilidade e aceitação do sistema Nota Conforme, tendo como base para adaptação o

modelo *Technology Acceptance Model* (TAM). Esse modelo, proposto por Fred Davis (1989), fundamenta-se em duas dimensões centrais que influenciam a aceitação de novas tecnologias: Percepções de Utilidade e Percepções de Facilidade de Uso. Além dessas dimensões, o modelo também considera a intenção comportamental, ou seja, a predisposição do usuário em utilizar a tecnologia no futuro (DAVIS; BAGOZZI; WARSHAW, 1989, p. 985). Essas dimensões constituem um indicativo importante para mensurar o potencial de aceitação e uso continuado do sistema.

As questões desta etapa foram formuladas para captar as percepções dos participantes em relação a essas três dimensões, permitindo uma análise estruturada do potencial de aceitação do sistema no ambiente da administração tributária. As respostas foram obtidas com base em uma escala de avaliação, conforme apresentado na Tabela 23, que possibilitou mensurar o grau de concordância dos participantes em relação às afirmações propostas.

Tabela 23 – Escala das respostas às questões da avaliação do perfil dos participantes.

ESCALA	QPP1 e QPP2	DEMAIS QUESTÕES
1	Nenhuma experiência ou familiaridade	Discordo Totalmente
2	Pouca experiência ou familiaridade	Discordo Parcialmente
3	Experiência ou familiaridade moderada	Neutro
4	Boa experiência ou familiaridade	Concordo parcialmente
5	Muita experiência ou domínio do tema	Concordo Totalmente

Fonte: Elaborado pelo autor

A coleta de dados referente à avaliação de viabilidade e aceitação do sistema Nota Conforme foi conduzida por meio de um questionário *online*, elaborado na plataforma Microsoft *Forms*. O instrumento utilizado para a pesquisa encontra-se disponível no APÊNDICE D deste trabalho. Para garantir que os participantes tivessem compreensão adequada sobre o sistema, foi realizada uma apresentação introdutória em fevereiro de 2025 para contextualizar os participantes quanto às funcionalidades, resultados e aplicações práticas do sistema.

Adicionalmente, o formulário foi disponibilizado com um vídeo explicativo (disponível em: https://youtu.be/eCXR6xzKwEg), o qual apresentou uma visão geral do funcionamento do Nota Conforme e exemplos de casos de uso. Essa estratégia visou assegurar que os participantes estivessem devidamente informados sobre a proposta do sistema antes de registrarem suas percepções.

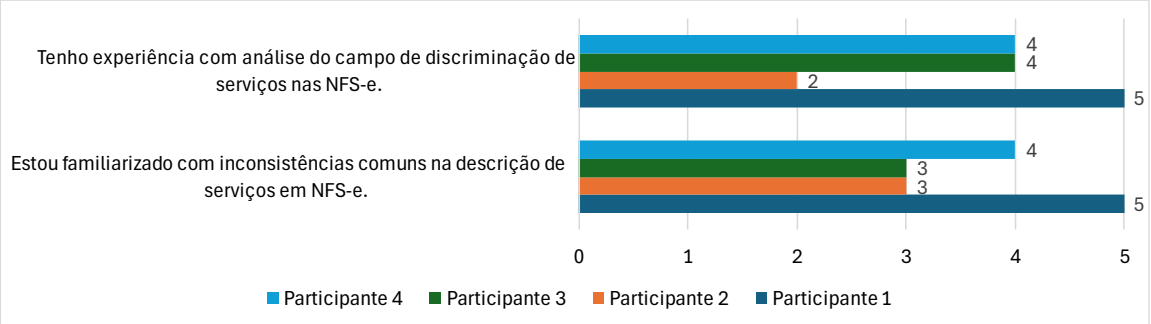
6.6.1. Perfil dos participantes

Nielsen (2000) defende que os testes de usabilidade com os melhores resultados são aqueles realizados com, no máximo, cinco usuários. Em termos de custo-benefício, o autor afirma que testes com três a cinco usuários são suficientes e oferecem resultados proporcionalmente ideais. Embora a presente avaliação não configure um teste de usabilidade, mas sim uma análise da percepção dos potenciais usuários em relação à aceitação da tecnologia proposta, adotaram-se os princípios defendidos pelo autor para a definição da quantidade de participantes.

Dessa forma, a pesquisa foi realizada com a participação de quatro potenciais usuários: três auditores fiscais e um gestor de modernização da administração tributária. Para assegurar a consistência e a confiabilidade das informações obtidas, o questionário foi respondido por profissionais com ampla e relevante experiência na área tributária, cujo tempo de atuação varia entre 5 e 11 anos, sendo que três deles possuem mais de nove anos de atuação.

A análise dos resultados referente as experiências com a análise de NFS-e revela níveis variados de familiaridade entre os participantes, conforme apresentados na Figura 36. Apesar dessa variação, constatou-se que todos os participantes possuem alguma experiência na análise do campo discriminação da NFS-e, bem como na competência na detecção e compreensão das inconsistências correlatadas. Esse cenário sugere a presença de um conhecimento técnico prévio, ainda que heterogêneo, habilitando os participantes a contribuírem de forma qualificada na pesquisa de viabilidade do Nota Conforme.

Figura 36 – Experiência com análise de NFS-e (QPP1)



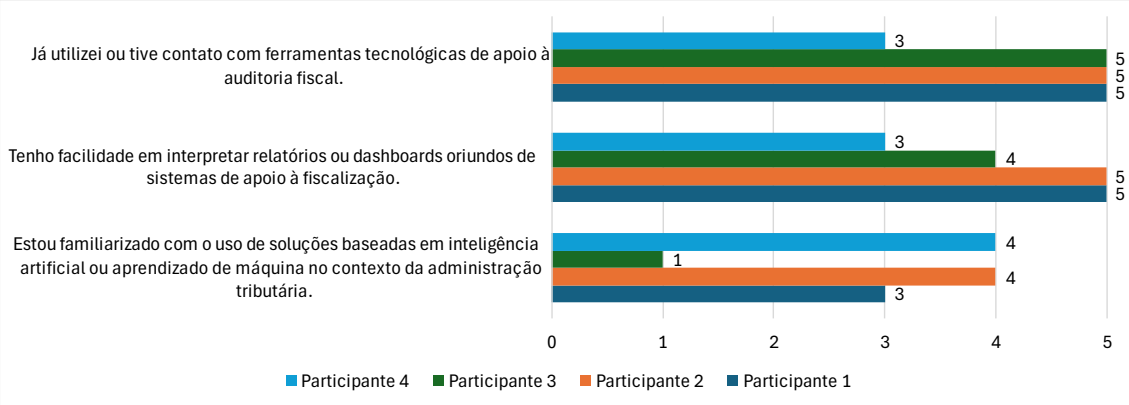
Fonte: Elaborado pelo autor

Conforme os dados apresentados na Figura 37, os participantes demonstraram familiaridade com tecnologias aplicadas à fiscalização tributária. Observou-se unanimidade

quanto ao uso prévio de ferramentas tecnológicas de apoio à auditoria fiscal, bem como a facilidade na interpretação de relatórios e dashboards provenientes desses sistemas.

No entanto, quando se trata do uso de soluções baseadas em inteligência artificial ou aprendizado de máquina no contexto da administração tributária, identificou-se um participante sem qualquer experiência ou familiaridade com essas tecnologias. Essa lacuna pode indicar uma oportunidade para a demonstração do potencial dessas técnicas através do Nota Conforme, contribuindo para a difusão de práticas inovadoras e o fortalecimento da cultura analítica no âmbito da fiscalização tributária.

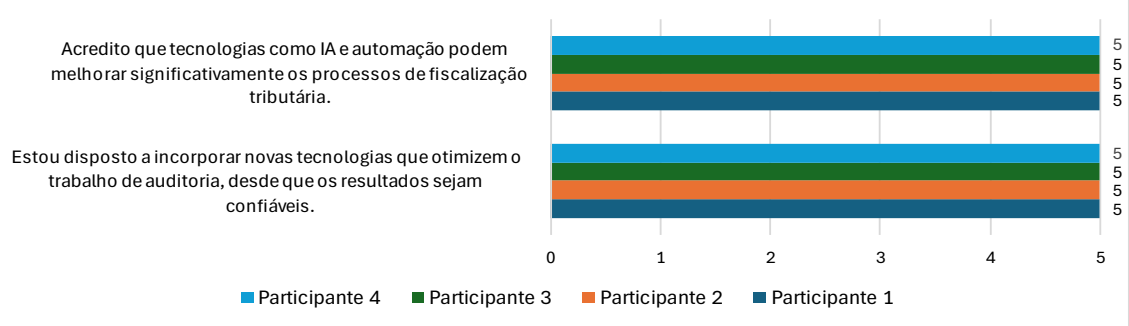
Figura 37 - Familiaridade com tecnologias aplicadas à fiscalização (QPP2)



Fonte: Elaborado pelo autor

Já no que se refere à percepção sobre inovação tecnológica no ambiente de trabalho, os dados revelam que os participantes são amplamente favoráveis, conforme Figura 38. Todos concordaram fortemente no potencial da aplicação de tecnologias como inteligência artificial e automação para melhoria dos processos de fiscalização tributária, bem como a disposição para incorporá-las, ao apresentarem resultados confiáveis.

Figura 38 - Percepção sobre inovação tecnológica no trabalho (QPP3)



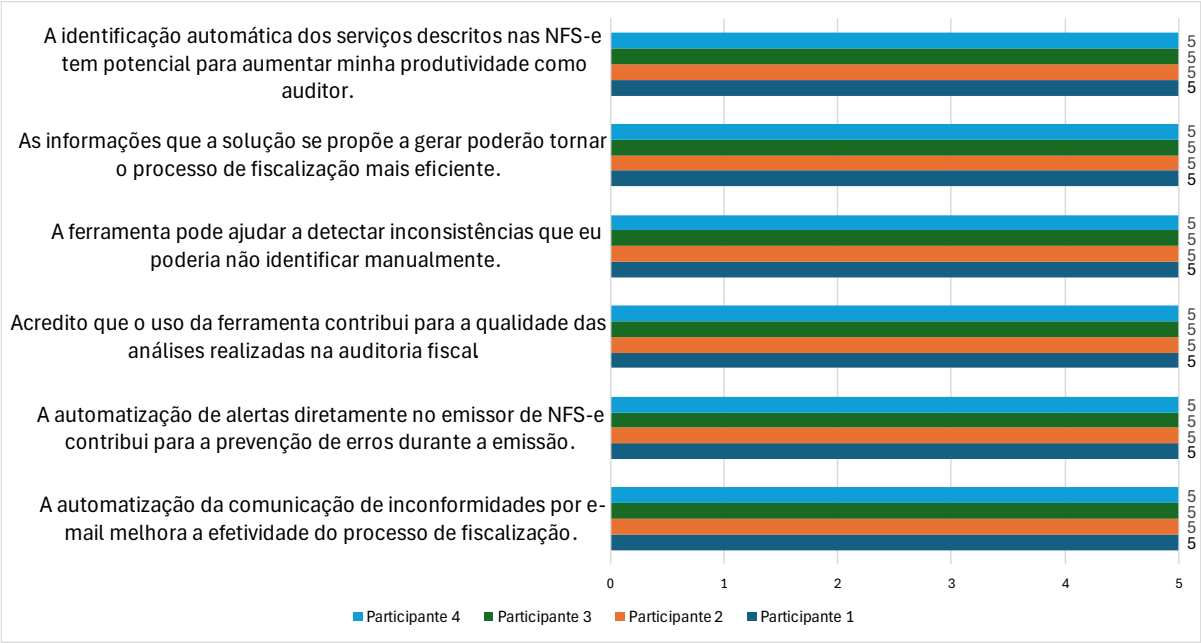
Fonte: Elaborado pelo autor

Esses resultados indicam não somente uma aceitação consciente da transformação digital no âmbito da auditoria fiscal, mas também uma abertura para adoção de soluções tecnológicas, representando um ambiente propício à implementação de iniciativas inovadoras, como o Nota Conforme. Além disso, os perfis dos participantes sugerem que as opiniões expressas por eles são tecnicamente fundamentadas e representam adequadamente o público-alvo da solução avaliada.

6.6.2. Percepções quanto a utilidade do Nota Conforme

A dimensão de utilidade percebida refere-se à avaliação se os usuários acreditam que o uso da tecnologia proposta melhora o desempenho do trabalho no contexto organizacional (DAVIS; BAGOZZI; WARSHAW, 1989, p. 985). Os resultados evidenciam concordância com as afirmações relacionadas à utilidade do sistema. Todos os participantes consideraram que a identificação automática dos serviços descritos nas NFS-e contribui para a qualidade das análises fiscais, tem potencial para aumentar a produtividade e tornar os processos de fiscalização mais eficientes. Também houve concordância quanto à capacidade do sistema de detectar inconsistências que poderiam passar despercebidas manualmente.

Figura 39 – Percepções quanto a utilidade do Nota Conforme (QTAM1)



Fonte: Elaborado pelo autor

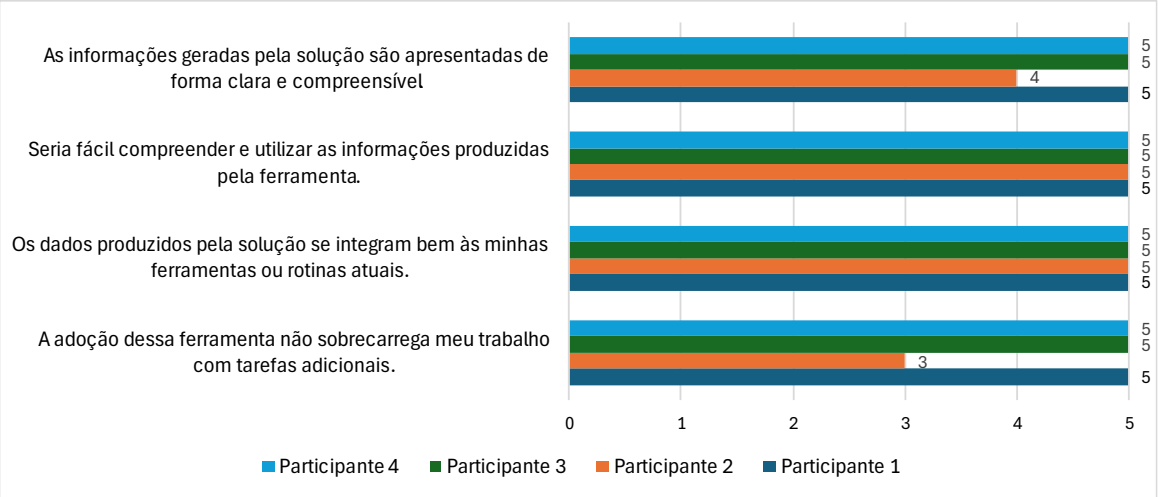
Além disso, os participantes afirmaram que a integração do Nota Conforme diretamente ao emissor de NFS-e, para alertar sobre possíveis inconsistências no momento da emissão, pode

contribuir significativamente para a prevenção de erros. Relataram também que o envio automatizado de relatórios de inconformidades, por meio da integração com ferramentas de disparo de e-mails, tende a aumentar a efetividade dos processos de fiscalização. Dessa forma, os resultados observados demonstram que o Nota Conforme é percebido como uma solução aplicável e relevante para o contexto prático da fiscalização tributária.

6.6.3. Percepções quanto a facilidade de uso do Nota Conforme

A dimensão de facilidade de uso percebida diz respeito ao grau em que o usuário acredita que a interação com a tecnologia será livre de esforço, ou seja, que seu uso será simples e intuitivo (DAVIS; BAGOZZI; WARSHAW, 1989). Os dados obtidos revelam uma percepção amplamente positiva quanto à usabilidade do sistema Nota Conforme, com destaque para os aspectos de clareza, compreensibilidade e integração das informações fornecidas. As respostas dos participantes indicam que os resultados gerados pela solução são considerados de fácil uso, tanto no que se refere à interpretação das informações quanto à sua aplicabilidade prática no contexto operacional das atividades de fiscalização.

Figura 40 - Percepções quanto a facilidade de uso do Nota Conforme (QTAM2)



Fonte: Elaborado pelo autor

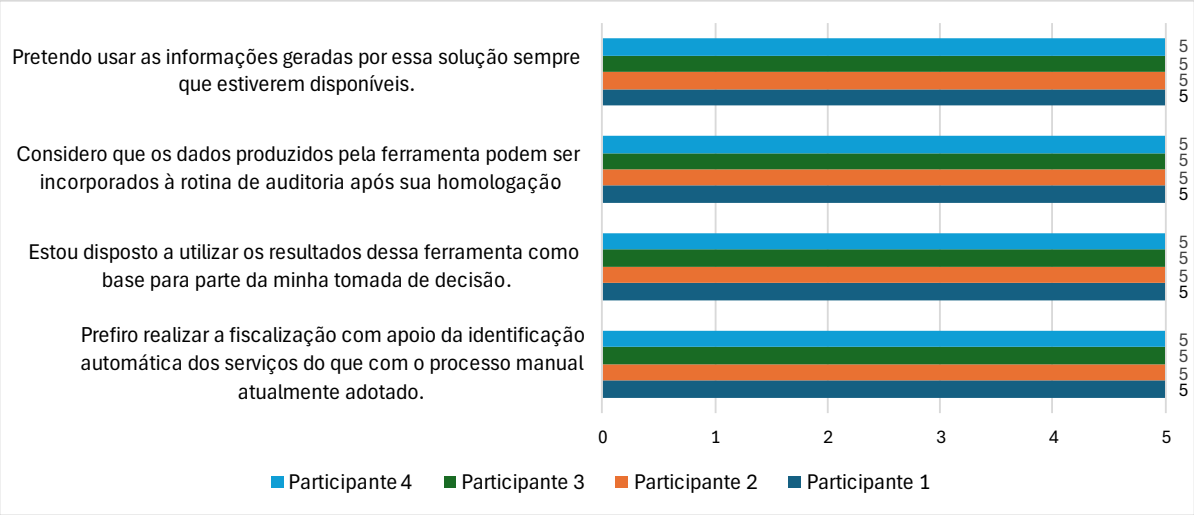
Além disso, os participantes relataram que a adoção do sistema não traria sobrecarga de trabalho, um fator relevante para a aceitação prática de novas tecnologias. Esses resultados indicam que a solução não apresenta barreiras significativas em termos de aprendizado ou esforço adicional. Isso sugere que o Nota Conforme apresenta bom potencial de aceitação, uma vez que combina facilidade de uso com integração às ferramentas e fluxos de trabalho existentes.

6.6.4. Percepções quanto ao possível uso do Nota Conforme no futuro

A intenção de uso futuro foi amplamente confirmada entre os participantes. Todos afirmaram que pretendem utilizar o Nota Conforme sempre que o sistema estiver disponível. Também houve consenso sobre a importância de incorporar os dados produzidos pelo sistema à rotina de auditoria.

Os participantes indicaram disposição em utilizar os resultados como base para a tomada de decisão. Por fim, todos demonstraram preferência pela fiscalização com apoio da ferramenta automatizada, em relação ao processo manual atualmente utilizado. Essas respostas apontam para uma alta predisposição à adoção efetiva da solução.

Figura 41 - Percepções quanto ao possível uso futuro do Nota Conforme (QTAM3)



Fonte: Elaborado pelo autor

A análise baseada no modelo TAM evidencia que o sistema Nota Conforme é bem aceito pelos potenciais usuários, sendo avaliado como: útil para o trabalho de auditoria fiscal; fácil de usar e de integrar às rotinas existentes; e com potencial para uso recorrente futuro. A aceitação entre profissionais experientes reforça o potencial do Nota Conforme como sistema inovador no apoio à fiscalização de NFS-e.

6.6.5. Aspectos positivos e sugestões de melhoria

Para oferecer aos participantes a oportunidade de expor aspectos positivos e negativos não contemplados no questionário baseado no modelo TAM, assim como sugerir melhorias para o Nota Conforme, foram disponibilizadas três questões com campo de resposta aberta. Os

participantes destacaram como aspectos positivos da utilização do Nota Conforme a agilidade na identificação de inconsistências nas descrições de serviços, a otimização do tempo de trabalho dos auditores da fiscalização, e a facilidade de uso da ferramenta, além da riqueza de conteúdo apresentado. Também foi ressaltado que o sistema permite à Secretaria de Finanças do Recife–PE acompanhar a evolução tecnológica e adaptar-se às exigências contemporâneas.

Um aspecto positivo relevante, apontado por um dos participantes, é que o sistema Nota Conforme permite ampliar a análise das NFS-e sem a necessidade de abertura de procedimentos fiscais, favorecendo a autorregularização por parte dos contribuintes. Essa característica contribui para o aumento da capacidade de monitoramento da administração tributária, possibilita a redução de custos operacionais e fortalece a cultura de cumprimento voluntário das obrigações fiscais.

Em relação aos aspectos negativos, destacou-se a necessidade de aprimorar a qualidade das respostas fornecidas pelo sistema, uma demanda natural no processo contínuo de evolução e refinamento de soluções baseadas em aprendizado de máquina. Como sugestão de melhoria, ressaltou-se a importância de uma curadoria mais detalhada dos dados utilizados no treinamento, visando aumentar a precisão das respostas geradas. A curadoria mencionada requer a disponibilização de mais recursos humanos e tempo para a realização das análises, o que poderá ser contemplado em iniciativas futuras.

Além disso, foi sugerido que o sistema incorporasse a funcionalidade de gerar relatórios de acompanhamento das notas fiscais retificadas pelos contribuintes, permitindo à administração tributária monitorar os efeitos da autorregularização. Tal funcionalidade está alinhada ao interesse deste pesquisador, a qual será incluída na seção de trabalhos futuros, por meio do desenvolvimento de uma plataforma *web*.

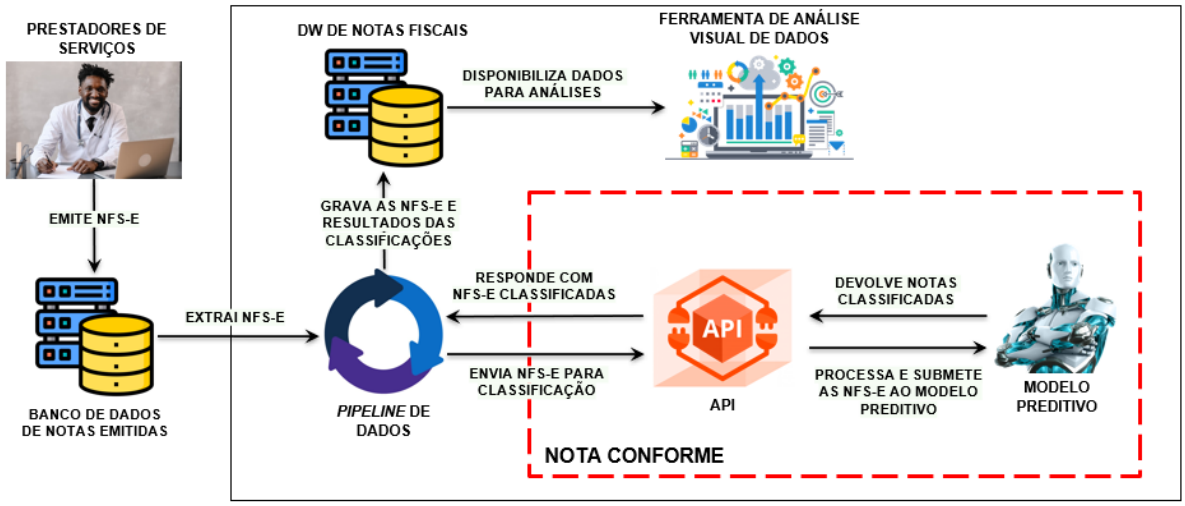
6.7. Estudo de caso do Nota Conforme

Com o objetivo de aprimorar os processos fiscais abordados nesta pesquisa e considerando os resultados promissores alcançados pelo Nota Conforme, os auditores propuseram três cenários para a implantação do sistema. Cada cenário contempla estratégias distintas de integração ao ambiente da Prefeitura do Recife–PE.

O primeiro cenário, ilustrado na Figura 42, o Nota Conforme, permitirá a integração com um *pipeline* de dados existente na Prefeitura do Recife–PE. Esse *pipeline* é responsável

pela extração de NFS-e da base de dados transacional, seu tratamento e carga no *Data Warehouse* (DW). Nesse contexto, as NFS-e extraídas pelo *pipeline* serão submetidas ao Nota Conforme através da API, que realizará as predições e retornará ao *pipeline* os códigos de serviços identificados nas descrições das NFS-e, para posterior gravação no DW.

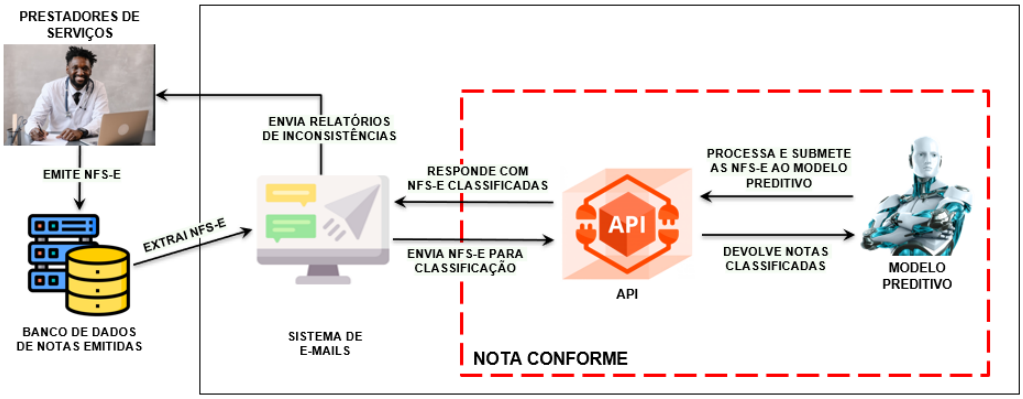
Figura 42 - Cena de uso do Nota Conforme para alimentar um DW.



Fonte: Elaborado pelo autor

Essa rotina pode ser programada conforme as necessidades das organizações, podendo ocorrer todos os dias, semanas ou meses. Na Prefeitura do Recife–PE, por exemplo, a rotina será executada semanalmente. Dessa forma, as NFS-e com os códigos de serviços identificados pelo Nota Conforme estarão disponíveis no DW, possibilitando que os auditores utilizem suas próprias ferramentas de visualização de dados para auxiliar na identificação de inconsistências nas NFS-e.

Figura 43 – Cena de uso do Nota Conforme integrado ao sistema de e-mails.



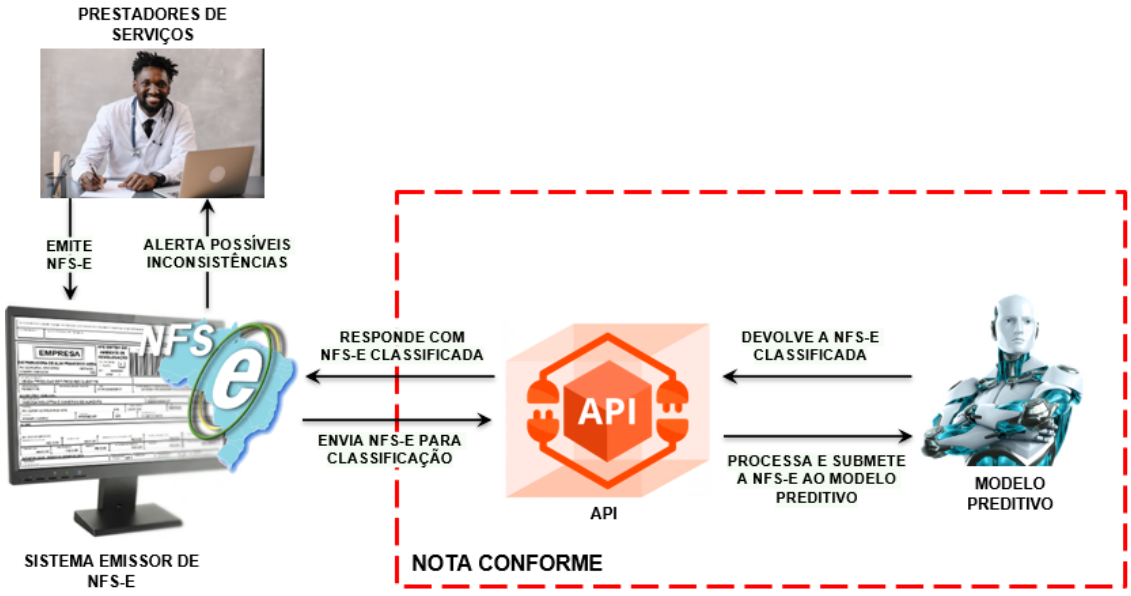
Fonte: Elaborado pelo autor

Na segunda cena de uso (Figura 43), o sistema Nota Conforme será integrado a uma plataforma de envio de e-mails, com o propósito de encaminhar automaticamente relatórios periódicos contendo as NFS-e que apresentarem inconsistências identificadas com base na análise dos serviços discriminados. Tal funcionalidade visa assegurar a notificação dos contribuintes acerca das irregularidades detectadas, possibilitando a correção nos prazos fiscais estabelecidos.

Os relatórios encaminhados poderão conter informações detalhadas, incluindo a identificação das notas, descrição das inconsistências e orientações para a regularização, promovendo, assim, maior controle e conformidade fiscal. Ademais, esta integração facilitará a comunicação entre os órgãos fiscalizadores e os responsáveis pela retificação, contribuindo para a mitigação de riscos relacionados a multas e sanções decorrentes de falhas na emissão.

No terceiro cenário, o Nota Conforme será integrado diretamente ao Sistema Emissor de NFS-e da Prefeitura do Recife–PE, conforme apresentado na Figura 44. Essa funcionalidade possibilitará que o emissor de NFS-e submeta, proativamente, a discriminação da NFS-e ao Nota Conforme logo após o seu preenchimento. Com o resultado da predição, será possível emitir os alertas de possíveis inconsistências para o usuário, antes mesmo da conclusão da emissão da NFS-e. Isso dará oportunidade para revisão da discriminação da NFS-e e correção de possíveis erros de maneira antecipada.

Figura 44 - Cena de uso do Nota Conforme integrado direto ao emissor de NFS-e.



Fonte: Elaborado pelo autor

É importante destacar, que esse tipo de funcionalidade não deve gerar nenhuma exigência de correção para o usuário. Conforme apresentado nas seções anteriores deste capítulo, existe uma margem de erro nas predições realizadas pelo modelo. Dessa forma, para evitar ocorrências de bloqueios indevidos que prejudiquem as emissões de NFS-e, seu uso deve ser exclusivo para geração dos alertas sobre possíveis inconsistências, permitindo que o usuário revise voluntariamente antes da confirmação da emissão da NFS-e.

Entretanto, esse cenário exige desenvolvimentos e melhorias no sistema emissor para integrá-lo ao Nota Conforme. Isso pode ser encarado como uma barreira, especialmente para prefeituras de cidades de menor porte, ao poderem não dispor de equipes de desenvolvimento disponível para desenvolver as melhorias ou não possuírem os códigos-fonte dos sistemas, por serem de propriedade de terceiros.

Portanto, o Nota Conforme pode ser implantado em diferentes contextos e órgãos. Por não exigir alterações nos sistemas de emissão de NFS-e, recomenda-se que sua implantação inicial seja via integração com DW por meio de *pipelines* de dados. Isso auxiliará o trabalho de fiscalização dos auditores, que poderão manter o desempenho do sistema em constante avaliação, bem como sugerir melhorias. Desse modo, o sistema poderá ser implantado, posteriormente, em ferramentas de disparo de e-mails ou integrada ao emissor de NFS-e com mais precisão e confiabilidade. Por fim, vale ressaltar que a solução será devidamente registrada no Instituto Nacional da Propriedade Industrial (INPI).

6.8. Respostas às questões de pesquisa

A resolução das questões de pesquisa envolveu uma revisão da literatura visando identificar técnicas e métodos de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (ML) utilizados para desenvolver soluções que resolvam problemas de classificação de textos, bem como a avaliação comparativa para escolha do modelo final. Trabalhos relacionados também foram explorados e discutidos a fim de conhecer técnicas aplicadas de maneira exitosa em problemas de classificação da área tributária. Além disso, as percepções dos potenciais usuários foram coletadas, a fim de avaliar a utilidade do sistema proposto.

6.8.1. Resposta à **QP01**

- **Questão de pesquisa:** Qual algoritmo de ML oferece o melhor equilíbrio entre precisão e consumo de recurso computacional na tarefa de identificar serviços em NFS-e?
- **Resolução:** Foi realizada uma avaliação comparativa entre diferentes algoritmos de classificação multirrótulo, para identificar a abordagem que proporcionasse o melhor equilíbrio entre desempenho preditivo e eficiência computacional. A partir dessa análise, foi selecionado o modelo baseado no algoritmo *Random Forest*, associado à técnica de representação textual *Term Frequency-Inverse Document Frequency* (TF-IDF). Esse modelo apresentou os melhores resultados na tarefa de classificação de serviços em NFS-e, alcançando um *F1-score* de 0,9969 na primeira avaliação e 0,945 na segunda, demonstrando elevada precisão e consistência.
- **Onde:** O processo de desenvolvimento do modelo é descrito no Capítulo 4, e os resultados obtidos são discutidos no Capítulo 6.
- **Tecnologias:** Python e Scikit-learn.

6.8.2. Resposta à **QP02**

- **Questão de pesquisa:** A automatização da identificação de serviços em NFS-e, utilizando ML, é útil e se integra com facilidade às rotinas e processos de fiscalização realizados pelos auditores?
- **Resolução:** O sistema Nota Conforme foi desenvolvido com foco na integração prática ao ambiente institucional da Prefeitura do Recife–PE, considerando aspectos como interoperabilidade com sistemas internos e usabilidade por parte dos auditores fiscais. Em testes realizados, a solução demonstrou desempenho promissor, com baixa latência de resposta, sendo avaliada positivamente quanto à sua viabilidade e aceitação. Os potenciais usuários destacaram a utilidade do sistema, sua facilidade de uso e a aderência às rotinas já estabelecidas de fiscalização, além de indicarem intenção de utilizá-la em futuras atividades.
- **Onde:** O sistema é detalhado no Capítulo 5, e os resultados de sua avaliação apresentados e discutidos no Capítulo 6.
- **Tecnologias:** Python e FastAPI.

6.9. Considerações finais

Este capítulo se propôs a detalhar os resultados desta pesquisa, a fim de avaliar a viabilidade de implantação do Nota Conforme em ambientes reais. Para isso, inicialmente, foram apresentados os resultados das duas avaliações comparativas que fundamentaram a escolha do algoritmo *Random Forest*. Os dados obtidos nessas etapas evidenciaram a capacidade do modelo em identificar corretamente os serviços descritos nas NFS-e, destacando seu desempenho superior em relação às demais abordagens testadas. Na seção seguinte, o desempenho do sistema foi comparado ao de trabalhos similares, evidenciando que os resultados obtidos pelo modelo apresentam desempenho compatível e encontram-se próximos aos reportados na literatura, demonstrando a efetividade da abordagem adotada.

Após avaliação do modelo preditivo, foi realizada uma avaliação na API, visando assegurar seu bom funcionamento para validar sua integração aos sistemas internos da Prefeitura do Recife–PE. Em seguida, foram discutidos os resultados da avaliação de viabilidade e aceitação do sistema Nota Conforme, com base na percepção dos potenciais usuários. Essa etapa permitiu analisar aspectos relacionados à utilidade, facilidade de uso e aplicabilidade prática da solução, contribuindo para uma compreensão mais ampla de seu potencial de adoção no contexto proposto.

Por fim, as questões de pesquisa foram discutidas, apresentando as respostas encontradas por esse estudo para cada uma delas. No capítulo seguinte, será apresentado um resumo dos principais achados deste estudo, juntamente com suas contribuições e as expectativas de aprimoramento em pesquisas futuras.

7. CONSIDERAÇÕES FINAIS

A presente pesquisa investigou métodos e técnicas de Inteligência Artificial (IA) que possibilitem o desenvolvimento de ferramentas para otimizar os processos dos órgãos de controle tributários, visando mitigar prejuízos aos cofres públicos causados pela prática de sonegação fiscal. Para isso, este estudo desenvolveu o Nota Conforme, um sistema composto por um modelo preditivo para identificar os serviços descritos na Nota Fiscais de Serviços Eletrônica (NFS-e) e uma *Application Programming Interface* (API) para implantação e integração com sistemas internos dos órgãos de controle. A API, foi desenvolvida na linguagem de programação Python com utilização do *framework* FastAPI.

O modelo desenvolvido foi selecionado a partir de uma avaliação comparativa envolvendo sete algoritmos de classificação. O modelo final adotou o algoritmo *Random Forest*, em conjunto com a representação textual baseada no método *Term Frequency-Inverse Document Frequency* (TF-IDF), combinação que demonstrou elevada eficácia na tarefa de classificação dos serviços descritos em NFS-e. Essa escolha foi respaldada pelos resultados obtidos em duas avaliações distintas, nas quais o desempenho do modelo destacou-se em termos de acurácia e consistência. Na primeira fase de avaliação, o modelo obteve desempenho promissor, resultando em uma acurácia de 0,993, precisão de 0,9978, *recall* de 0,9960 e *F1-Score* de 0,9969.

Já na segunda avaliação, realizada em 500 NFS-e selecionadas pelos Auditores do Tesouro Municipal da Prefeitura do Recife–PE, o modelo obteve precisão de 0,941, *recall* de 0,952 e *F1-Score* de 0,945. Apesar de estarem ligeiramente abaixo, os resultados são promissores e permanecem nos limites aceitáveis. Esses resultados evidenciam a robustez do modelo na identificação precisa dos serviços prestados, destacando sua capacidade de classificação automática com alta confiabilidade. Tal desempenho reforça o potencial da abordagem para aplicação em contextos reais, onde a correta categorização dos serviços é fundamental.

Os testes realizados na API demonstraram sua viabilidade para implantação, mesmo em ambientes com grandes volumes de dados. O sistema apresentou uma latência de 6,9 milissegundos no processamento e na devolução do resultado da predição para cada NFS-e. Além disso, a funcionalidade de campos dinâmicos nos atributos de entradas e respostas das

requisições da API funcionou corretamente, sem impactar o desempenho geral do sistema, permitindo sua implementação em diferentes contextos.

Na prática, o Nota Conforme possibilita a identificação automatizada dos serviços prestados, conforme são descritos na discriminação da NFS-e. Assim, espera-se que os processos de fiscalizações sejam otimizados, pois os auditores poderão utilizar ferramentas de análise visual de dados para realizar cruzamentos dos dados das NFS-e com os resultados das predições para identificar eventuais inconsistências tributárias. Além disso, a solução poderá ser integrada a sistemas internos por meio da API, permitindo a automatização de outras tarefas como, por exemplo, o envio de alertas proativos aos contribuintes sobre possíveis inconsistências em NFS-e, para poderem realizar as devidas correções.

Entretanto, é importante destacar alguns desafios e dificuldades enfrentados. Alguns serviços são prestados com menor frequência e, por consequência, possuem poucos exemplos em base de dados, dificultando o aprendizado do modelo. Outro desafio decorre pelo fato das discriminações genéricas apresentarem uma grande variabilidade de termos, uma vez que podem conter qualquer tipo de informação. Isso torna a predição mais difícil, aumentando a probabilidade do modelo cometer erros do tipo Falsos Negativos (FN). Além disso, o modelo poderá apresentar desempenho inferior em outros municípios, pois foi treinado exclusivamente com dados do emissor de NFS-e da Prefeitura do Recife–PE, o que pode limitar sua generalização para diferentes contextos e características locais.

Apesar dos desafios enfrentados, o sistema Nota Conforme apresentou desempenho satisfatório, configurando-se como uma solução promissora para a classificação automática de serviços descritos em NFS-e. Ademais, sua aceitação e viabilidade no contexto da Prefeitura do Recife–PE foram evidenciadas por meio da percepção dos usuários potenciais, que reconheceram sua utilidade, destacaram a facilidade de integração às rotinas atuais de fiscalização e demonstraram intenção de utilizá-la futuramente.

7.1. Contribuições

Esta pesquisa foi desenvolvida a fim de contribuir e aprimorar os processos da Administração Tributária. Entre as principais contribuições, destacam-se as seguintes soluções:

- Desenvolvimento de um modelo preditivo para identificação de serviço a partir do processamento da discriminação da NFS-e.

- Desenvolvimento de API para implantação e integração do modelo com sistemas internos dos órgãos de controle.

7.2. Limitações

Apesar dos resultados promissores obtidos com a aplicação do modelo de classificação multirrótulo desenvolvido no escopo do sistema Nota Conforme, é fundamental reconhecer as limitações que restringem a generalização e aplicabilidade imediata da solução.

A principal limitação diz respeito à restrição setorial da base de dados utilizada para treinamento e validação do modelo, composta exclusivamente por NFS-e do setor de saúde. Esse domínio possui características linguísticas e estruturais específicas, com vocabulário técnico mais delimitado e maior padronização nas descrições, o que pode ter contribuído positivamente para os bons resultados observados em termos de acurácia, *F1-score* e *Hamming Loss*.

No entanto, ao considerar a diversidade de setores econômicos presentes na arrecadação municipal, torna-se evidente que outras áreas como, por exemplo, construção civil, serviços de informática, consultoria, estética ou manutenção também apresentam vocabulário mais genérico, descrições informais e maior ambiguidade semântica. Essa diversidade pode comprometer o desempenho do modelo atual, caso seja aplicado diretamente sem readequações. Assim, a capacidade de generalização do classificador para outros domínios permanece uma hipótese não testada e constitui uma limitação metodológica relevante.

Além disso, a restrição às NFS-e do município do Recife–PE impõe um recorte geográfico relevante. Essa delimitação compromete a capacidade de generalização do modelo preditivo para outros contextos regionais, nos quais aspectos como a terminologia empregada, estrutura dos dados e padrões de emissão fiscal podem diferir significativamente. Como consequência, a aplicabilidade da solução em outros municípios pode ser limitada.

Adicionalmente, o processo de rotulagem dos dados contou com a participação de auditores especializados no setor de saúde, o que, embora positivo, introduz um viés de escopo ao processo de categorização. A dependência do conhecimento de domínio específico também representa um desafio para a escalabilidade do sistema para múltiplas áreas.

Embora o processamento textual empregado utilizado seja eficiente em tarefas de vetorização, ele pode apresentar limitações na captura de nuances semânticas, especialmente

em descrições curtas, ambíguas ou que utilizam linguagem técnica específica. Além disso, a escolha por modelos clássicos de aprendizado de máquina, como regressão logística e árvores de decisão, embora sejam reconhecidos por sua interpretabilidade e bom desempenho em diversos cenários, possuem limitações quanto à capacidade de capturar relações complexas nos dados. A ausência de modelos mais avançados, como redes neurais profundas ou *Large Language Models* (LLM), representa uma oportunidade não explorada.

Outra limitação a ser destacada refere-se à ausência de mecanismos de explicabilidade algorítmica no modelo disponibilizado via API. Em aplicações de Inteligência Artificial (IA) no setor público, a capacidade de justificar ou interpretar a saída do modelo é essencial para assegurar transparência e responsabilização, especialmente em contextos regulatórios e fiscais. Por fim, do ponto de vista da pesquisa qualitativa realizada com potenciais usuários, ainda que os resultados tenham sido positivos em relação à utilidade e aceitação do sistema, a amostra é limitada ao contexto da Prefeitura do Recife-PE. A replicação desse estudo de aceitação em outros contextos institucionais poderá fornecer insumos adicionais para validar a viabilidade da ferramenta em escala mais ampla.

7.3. Trabalhos futuros

Diante das discussões apresentadas, identificam-se diversas oportunidades para a continuidade e aprimoramento deste estudo em pesquisas futuras, tais como:

- Expandir o escodo para incluir os serviços dos demais setores econômicos, já que a versão atual está limitada aos serviços do setor de saúde.
- Ampliar a amostra de treinamento, incorporando mais exemplos reais das classes menos frequentes.
- Expandir a curadoria da base de dados utilizada no treinamento, visando aumentar a qualidade e representatividade dos dados.
- Incorporar técnicas de explicabilidade algorítmica (como LIME, SHAP ou análise de importância de termos) na API do sistema, oferecendo aos auditores maior transparência na tomada de decisão;
- Aprimorar o modelo com o uso de LLM, visando alcançar maior capacidade de generalização e desempenho.
- Avaliar o impacto legal e ético da automação de decisões fiscais com IA considerando aspectos como viés, erro e responsabilização;

- Ampliar a pesquisa qualitativa com usuários de outras prefeituras, validando percepções de utilidade, facilidade e confiabilidade em contextos institucionais diversos.
- Desenvolver uma plataforma *web* para visualização e gerenciamento das predições geradas pelo modelo preditivo, permitindo o controle das NFS-e e dos contribuintes fiscalizados, além de oferecer suporte aos auditores por meio de relatórios de acompanhamento.
- Integrar técnicas de aprendizado por reforço, permitindo que o sistema aprenda a maximizar a precisão de suas predições ao longo do tempo, com base no *feedback* das ações anteriores.
- Ampliar o escopo, incorporando novos atributos da NFS-e, que possibilitem a apuração automática dos tributos evadidos.

7.4. Considerações finais

Dessa forma, com a implementação da solução, espera-se que os órgãos de controle e arrecadação possam aperfeiçoar seus processos de fiscalização, auditorias e prevenção à evasão fiscal. Assim, a sua utilização permitirá uma análise mais assertiva dos dados, possibilitando a identificação precoce de irregularidades e a implementação de medidas corretivas de forma mais ágil. Com isso, será possível melhorar a arrecadação tributária, reduzir fraudes fiscais e, consequentemente, contribuir para o desenvolvimento econômico sustentável e para a justiça fiscal no município.

8. REFERÊNCIAS

ABRASF – Associação Brasileira das Secretarias de Finanças das Capitais. **NFS-e**: Modelo Conceitual. Versão 1, Brasília, 29 dez. 2008. Disponível em: https://abrasf.org.br/biblioteca/arquivos-publicos/nfs-e/versao-1-00. Acesso em: 10 out. 2024.

BRASIL. **Decreto n.º 6.022, de 22 de janeiro de 2007.** Institui o Sistema Público de Escrituração Digital - SPED. Diário Oficial da União: seção 1, Brasília, DF, 23 jan. 2007. Disponível em: https://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2007/Decreto/D6022.htm. Acesso em: 10 nov. 2024.

BRASIL. **Lei n.º 14.822, de 22 de janeiro de 2024**. Estima a receita e fixa a despesa da União para o exercício financeiro de 2024. Diário Oficial da União: seção 1, Brasília, DF, p. 1, 23 jan. 2024. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/L14822.htm. Acesso em: 19 nov. 2024.

BRASIL. **Lei nº 4.729, de 14 de julho de 1965**. m,Define o crime de sonegação fiscal e dá outras providências. Brasília, 1965. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/1950-1969/l4729.htm. Acesso em: 20 jul. 2024.

BREIMAN, Leo. Consistency for a Simple Model of Random Forests. **Technical Report 670**, Department of Statistics, University of California, Berkeley, 2004.

BREIMAN, Leo. Random Forests. **Machine Learning**, [*s. l.*], v. 45, n. 1, p. 5-32, 2001.

BRUCE, Peter; BRUCE, Andrew. **Estatística Prática para Cientista de Dados:** 50 Conceitos Essenciais. Tradução: Luciana Ferraz. Rio de Janeiro: Alta Books, 2019. 320 p.

C4 MODEL. **C4 Model for visualising software architecture**. Disponível em: https://c4model.com/. Acesso em: 12 ago. 2024.

CASELI, H.M.; NUNES, M.G.V. (org.). **Processamento de Linguagem Natural**: Conceitos, Técnicas e Aplicações em Português. 2 ed. BPLN, 2024. Disponível em: https://brasileiraspln.com/livro-pln. Acesso em: 01 jul. 2024.

CHAI, C. P. Comparison of text preprocessing methods. **Natural language engineering**, v. 29, n. 3, p. 509–553, 2023. DOI: 10.1017/S1351324922000213. Disponível em: https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/comparison-of-text-preprocessing-methods/43A20821D65F1C0C4366B126FC794AE3. Acesso em: 10 out. 2024.

DAVIS, F. D.; BAGOZZI, R. P.; WARSHAW, P. R. User acceptance of computer technology: a comparison of two theoretical models. *Management Science*, v. 35, n. 8, p. 982–1003, 1989. Disponível em: http://www.jstor.org/stable/2632151. Acesso em: 01 jul. 2025.

DE ANGELI NETO, Humberto; MARTINEZ, Antonio Lopo. Nota Fiscal de Serviços Eletrônica: uma análise dos impactos na arrecadação em municípios brasileiros. **Revista de Contabilidade e Organizações**, São Paulo, Brasil, v. 10, n. 26, p. 49–62, 2016. DOI: 10.11606/rco.v10i26.107117. Disponível em: https://www.revistas.usp.br/rco/article/view/107117. Acesso em: 15 nov. 2024.

DE ARAUJO NETO, Antonio Marinho. **O uso de processamento de linguagem natural para classificação de produtos no contexto de notas fiscais eletrônicas**. 2021. 25 p.

Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade Estadual da Paraíba, Campina Grande, 2021.

DE CARVALHO, F. A. A. Gestão Fiscal Estratégica por Meio do Uso da Inteligência Artificial e o Princípio da Eficiência Administrativa: análise aplicada no âmbito do Município de João Pessoa. **Revista da Procuradoria Geral do Município de João Pessoa**, João Pessoal, Brasil, v. 9, 2024. DOI: 10.71144/2966-4977.9.2024.16. Disponível em: https://revistapgmjp.com.br/index.php/ojs/article/view/16. Acesso em: 7 jan. 2025.

DIAS, M.; BECKER, K. Identificação de Candidatos à Fiscalização por Evasão do Tributo ISS. In: Symposium on Knowledge Discovery, Mining and Learning, 5, 2017, Uberlândia. **Anais [...].** Uberlândia: Sociedade Brasileira de Computação, 2017.

DOS ANJOS, Pedro Germano; PINHEIRO, Marcelo Teles Silva. A implementação da Inteligência Artificial (IA) na fiscalização tributária: inovações disruptivas para eficiência na arrecadação do IPTU. **Revista Tributária e de Finanças Públicas**, [s. *l.*], v. 159, 16 jun. 2024. Disponível em: https://rtrib.abdt.org.br/index.php/rtfp/article/view/719. Acesso em: 30 nov. 2024.

FACELI, Katti. et al. **Inteligência artificial**: uma abordagem de aprendizado de máquina. 2 ed. Rio de Janeiro: LTC, 2023. 400 p.

FERILLI, S.; ESPOSITO, F.; GRIECO, D. Automatic learning of linguistic resources for stopword removal and stemming from text. **Procedia computer science**, v. 38, p. 116–123, 2014. DOI: 10.1016/j.procs.2014.10.019. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050914013799. Acesso em: 03 jul. 2024.

FNS - Fundo Nacional de Saúde. Conheça os valores para apresentação de propostas ao MS em 2023. Brasília, 2023. Disponível em: https://portalfns.saude.gov.br/conheca-os-valores-para-apresentacao-de-propostas-ao-ms-em-2023/. Acesso em: 18 fev. 2025.

GÉRON, Aurélien. **Mãos à Obra Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow: Conceitos, Ferramentas e Técnicas Para a Construção de Sistemas Inteligentes.** Tradução: Cibelle Ravaglia. Rio de Janeiro: Alta Books, 2021. 640 p.

GERON, Cecília M. S. et al. SPED – Sistema Público de Escrituração Digital: Percepção dos contribuintes em relação os impactos de sua adoção. **Revista de Educação e Pesquisa em Contabilidade (REPeC)**, [*S. l.*], v. 5, n. 2, p. 44–67, 2011. DOI: 10.17524/repec.v5i2.343. Disponível em: https://www.repec.org.br/repec/article/view/343. Acesso em: 18 fev. 2025.

GOMES, Wesckley Faria. **Análise exploratória e experimental de aplicações de inteligência artificial para classificação de descrições incongruentes em compras na área de saúde pública**. 64 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Sergipe, São Cristóvão, 2023.

HUYEN, Chip. **Projetando Sistemas de Machine Learning: processo interativo para aplicações prontas para produção**. Tradução de Cibelle Ravaglia. 1. ed. Rio de Janeiro: Alta Books, 2023. 384 p.

IBPT - INSTITUTO BRASILEIRO DE PLANEJAMENTO E TRIBUTAÇÃO. **Estudo sobre sonegação fiscal das empresas brasileiras.** 2023. Disponível em: https://ibpt.org.br/estudo-sobre-sonegacao-fiscal-das-empresas-brasileiras/. Acesso em: 04 jan. 2025.

JOBLIB DEVELOPERS. **Joblib documentation**. Joblib, 2024. Disponível em: https://joblib.readthedocs.io/en/stable/. Acesso em: 14 nov. 2024.

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models**. 3. ed. Manuscrito online, 2025. Disponível em: https://web.stanford.edu/~jurafsky/slp3. Acesso em: 16 jan. 2025.

JWT.io. JSON Web Tokens. Disponível em: https://jwt.io/. Acesso em: 24 nov. 2024.

KALINOWSKI, M. et al. **Engenharia de Software para Ciência de Dados**: Um guia de boas práticas com ênfase na construção de sistemas de machine learning. São Paulo: Casa do Código, 2023.

KANG, Y. et al. Natural language processing (NLP) in management research: A literature review. **Journal of management analytics**, v. 7, n. 2, p. 139–172, 2020. DOI: 10.1080/23270012.2020.1756939. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/23270012.2020.1756939. Acesso em: 15 jun. 2024.

LINS NETO, José Correia. **Audita-NFSe: sistema auxiliar de auditoria em notas fiscais de serviços eletrônicas**. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Pernambuco, Recife, 2021.

LIU, Cai-zhi. et al. Research of text classification based on improved TF-IDF algorithm. In: 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE). Lanzhou, China. **Anais[...]** IEEE, 2018.

MACEDO, C.; DINIZ FILHO, J. W. de F. SONEGAÇÃO FISCAL: Uma análise dos seus Efeitos na Economia Brasileira. **Revista de Auditoria Governança e Contabilidade**, Monte Carmelo, Brasil, v. 7, n. 31, 2019. Disponível em: https://revistas.fucamp.edu.br/index.php/ragc/article/view/1889. Acesso: 16 out. 2024.

NIELSEN, Jakob. Why you only need to test with 5 users. Nielsen Norman Group, 18 mar. 2000. Disponível em: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/. Acesso em: 10 set. 2025.

PANDAS DEVELOPMENT TEAM. Pandas documentation. Disponível em: https://pandas.pydata.org/docs/reference/index.html. Acesso em: 20 dez. 2024.

PARMAR, A.; KATARIYA, R.; PATEL, V. A review on random forest: An ensemble classifier. In: International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. **Springer International Publishing**, 2019. p. 758–763. DOI: 10.1007/978-3-030-03146-6_86. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-03146-6_86. Acesso: 25 jul. 2024.

PINHEIRO, Geraldo José; CUNHA, Luís R. Silva. A importância da auditoria na detecção de fraudes. **Contabilidade Vista & Revista**, [*S. l.*], vol. 14, núm. 1, abril, 2003, pp. 31-47. Universidade Federal de Minas Gerais. Minas Gerais, Brasil. Disponível em: https://revistas.face.ufmg.br/index.php/contabilidadevistaerevista/article/view/210. Acesso em: 30 nov. 2024.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. Metodologia do trabalho científico [recurso eletrônico]: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Feevale, 2013.

PYTHON SOFTWARE FOUNDATION. Python 3.13 documentation. Disponível em: https://docs.python.org/3.13/index.html. Acesso em: 20 dez. 2024.

RAMÍREZ, Sebastián. **FastAPI**. Disponível em: https://fastapi.tiangolo.com/. Acesso em: 15 nov. 2024.

RECIFE. **Manual de Utilização do Web Service da NFS-e**: Modelo Nacional, versão 1.1. 2017. Disponível em: https://nfse.recife.pe.gov.br/arquivos/WsNFSeNacional.pdf. Acesso em: 10 out. 2024.

RECIFE. **NFSe - Nota Fiscal de Serviços Eletrônica**. Prefeitura do Recife, 2024a. Disponível em: https://nfse.recife.pe.gov.br. Acesso em: 20 jul. 2024.

RECIFE. **Nota Fiscal de Serviço Eletrônica (NFS-e):** Acesso ao Sistema Pessoa Jurídica. Prefeitura do Recife, 2024b. Disponível em: https://nfse.recife.pe.gov.br/arquivos/NFSe_PJ.pdf. Acesso em: 25 out. 2024.

RECIFE. **Decreto n.º 23.675, de 30 de maio de 2008**. Regulamenta a Lei 17.407 de 02 de janeiro de 2008, que institui a Nota Fiscal Eletrônica de Serviços, NFS-e e a Lei 17.408 de 20 de marco de 2008, que dispõe sobre a geração e utilização de créditos tributários para tomadores de serviço. Recife, 2008a. Disponível em: https://nfse.recife.pe.gov.br/Arquivos/Decreto_23675_2008.pdf. Acesso em: 20 jul. 2024.

RECIFE. **Decreto n.º 24.093 de 05 de novembro de 2008**. Regulamenta o preenchimento da Nota Fiscal de Serviços Eletrônica, NFS-e, instituída pela Lei nº. 17.407 de 02 de janeiro de 2008. Recife, 2008b. Disponível em: https://nfse.recife.pe.gov.br/Arquivos/Decreto_24093_2008.pdf. Acesso em: 20 jul. 2024.

RECIFE. **Lei n.º 17.407, de 2 de janeiro de 2008**. Institui a Nota Fiscal de Serviços Eletrônica. Recife, 2008c. Disponível em: https://nfse.recife.pe.gov.br/Arquivos/lei_17407_2008.pdf. Acesso em: 20 jul. 2024.

RUSSELL, Stuart; NORVIG, Peter. Inteligência artificial: uma abordagem moderna. Trad.: Daniel Vieira; Flávio Soares Corrêa da Silva. 4ª Ed. Rio de Janeiro: Grupo Editora Nacional, 2022. 1016 p.

SCIKIT-LEARN DEVELOPERS. **API Reference - scikit-learn 1.5.2 documentation**. Scikit-Learn, 2024. Disponível em: https://scikit-learn.org/stable/api/index.html. Acesso em: 10 nov. 2024.

SCORNET, Erwan; BIAU, Gérard; VERT, Jean-Philippe. **Consistency of random forests**. The Annals of Statistics, v. 43, n. 4, p. 1716–1741, ago. 2015. DOI: 10.1214/15-AOS1321. Disponível em: https://projecteuclid.org/journals/annals-of-statistics/volume-43/issue-4/Consistency-of-random-forests/10.1214/15-AOS1321.full. Acesso em: 10 jul. 2024.

SEBRAE. Quais as diferenças entre as notas fiscais: NF-e e NFS-e? Florianópolis, 2023. Disponível em: https://www.sebrae-sc.com.br/blog/diferencas-entre-as-notas-fiscais-nfe-e-nfse. Acesso em: 18 nov. 2024.

SHAH, A. A.; RANA, K. A review on supervised machine learning text categorization approaches. International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET). **IEEE**, Kottayam, India, 2018, p. 1-6, DOI: 10.1109/ICCSDET.2018.8821134. Disponível em: https://ieeexplore.ieee.org/document/8821134. Acesso em: 30 jul. 2024.

SILVA, A. F. DA et al. SPED – Public Digital Bookkeeping System: influence in the economic-financial results declared by companies. **Review of Business Management**, *[S. l.]*, v. 15, n. 48, p. 445–461, 2013. DOI: 10.7819/rbgn.v15i48.1330. Disponível em: https://rbgn.fecap.br/RBGN/article/view/1330. Acesso em: 9 jan. 2025.

SOARES, Glauco de Vasconcelos; CUNHA, Rodrigo. Predição de Irregularidade Fiscal dos Contribuintes do Tributo ISS. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD), 35., 2020, Evento Online. **Anais [...]**. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 223-228. ISSN 2763-8979. DOI: 10.5753/sbbd.2020.13645. Disponível em: https://sol.sbc.org.br/index.php/sbbd/article/view/13645. Acesso em: 23 jun. 2024.

SONG, Qing; LIU, Xiaoou; YANG, Lu. The random forest classifier applied in droplet fingerprint recognition. In: INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY (FSKD), 12, 2015, Zhangjiajie, China. **Anais [...]**, [*S. l.*]: IEEE, 2015. p. 722-726. DOI: 10.1109/FSKD.2015.7382031. Disponível em: https://ieeexplore.ieee.org/document/7382031. Acesso em: 02 jul. 2024.

UYSAL, A. K.; GUNAL, S. The impact of preprocessing on text classification. **Information Processing & Management**, v. 50, n. 1, p. 104–112, jan. 2014. DOI: 10.1016/j.ipm.2013.08.006. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0306457313000964. Acesso em: 15 nov. 2024.

VAJJALA, S. et al. **Practical natural language processing:** A comprehensive guide to building real-world NLP systems. Sebastopol, CA, USA: O'Reilly Media, 2020. 422 p.

WAZLAWICK, Raul Sidnei. Metodologia de pesquisa para ciência da computação. 3. ed. Rio de Janeiro: LTC, 2021.

APÊNDICES

APÊNDICE A – Artigo: Avaliação do desempenho de algoritmos de classificação na identificação de serviços em NFS-e.

Avaliação do desempenho de algoritmos de classificação na identificação de serviços em NFS-e

Tarcísio Paraiso Farias[1], **Thiago Souto Mendes**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA)
Programa de Pós-Graduação em Engenharia de Sistemas e Produtos

tarcisioparaiso@gmail.com, thiagosouto@ifba.edu.br

Abstract. *The Electronic Service Invoice (NFS-e) has proven to be a relevant tool for the modernization of municipal tax collection. However, fraud remains frequent, resulting in significant losses to public funds. This study proposes and experimentally evaluates the application of machine learning (ML) algorithms for the automatic classification of services described in NFS-e, aiming to support the detection of tax inconsistencies and the correct assignment of tax rates. The performances of seven ML algorithms were compared. The results indicate that the Random Forest, Decision Tree, Linear Support Vector, and LightGBM algorithms achieved the best performances. It is noteworthy that, although Random Forest and Decision Tree require more time and computational resources for training and prediction, Linear Support Vector and LightGBM showed competitive performance with significantly lower computational costs.*

Resumo. *A Nota Fiscal de Serviços Eletrônica (NFS-e) tem se mostrado uma ferramenta relevante para a modernização da arrecadação tributária municipal. No entanto, fraudes ainda são recorrentes, acarretando perdas significativas para os cofres públicos. Este trabalho propõe e avalia experimentalmente a aplicação de algoritmos de aprendizado de máquina (ML) para a classificação automática dos serviços descritos nas NFS-e, com o objetivo de apoiar a detecção de inconsistências fiscais e o correto enquadramento de alíquotas. Foram comparados os desempenhos de sete algoritmos de ML. Os resultados indicam que os algoritmos Random Forest, Decision Tree, Linear Support Vector e LightGBM obtiveram os melhores desempenhos. Destaca-se que, embora Random Forest e Decision Tree exijam mais tempo e recursos computacionais para treinamento e predição, o Linear Support Vector e o LightGBM apresentaram desempenhos competitivos com custos computacionais significativamente menores.*

1. Introdução

A Nota Fiscal de Serviços Eletrônica (NFS-e) surgiu com a finalidade de registrar operações de prestação de serviços [ABRASF 2008]. Sua implantação promove maior agilidade na arrecadação de tributos [Neto and Martinez 2016]. Entretanto, muitas empresas recorrem a mecanismos para bular as regras fiscais com intuito de pagar menos tributos [Dias and Becker 2017]. Isso causa prejuízos ao erário e compromete investimentos públicos em áreas fundamentais da sociedade, tornando a evasão fiscal um obstáculo para a administração pública [De Macedo and Diniz Filho 2019].

Nesse cenário, o monitoramento e a análise de dados fiscais possui a finalidade de fazer cumprir as obrigações fiscais [Pinheiro and Cunha 2009]. No contexto das NFS-e, o campo descritivo dos serviços prestados assume papel central nos processos de fiscalização, uma vez que contém informações detalhadas acerca da natureza das atividades executadas. Essas informações viabilizam o cruzamento entre os cruzamento dos serviços prestados e as alíquotas aplicadas. Entretanto, esse campo é preenchido em linguagem natural, o que dificulta o uso de técnicas tradicionais, como comandos em Linguagem de Consulta Estruturada (SQL) [Lins Neto 2021]. Por isso, são necessárias abordagens mais avançadas de análise textual. Além disso, de acordo com dados coletados pelos autores, no primeiro semestre de 2024 foi registrada uma média diária de 82 mil NFS-e emitidas no Recife. Esse volume evidencia a inviabilidade da auditoria manual, como já discutido em [Neto and Martinez 2016], reforçando a necessidade de soluções automatizadas para análise fiscal.

A implementação de soluções de Inteligência Artificial (IA) pode melhorar os processos de auditoria tributária [Dos Anjos and Pinheiro 2024]. Em razão disso, identificou-se a necessidade do uso de técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (ML) para automatizar a identificação de serviços descritos na NFS-e, auxiliando na detecção de fraudes por meio do cruzamento de serviços executados e alíquotas praticadas. Este artigo apresenta uma avaliação experimental de algoritmos de ML na classificação de serviços em NFS-e, abordando etapas de coleta, pré-processamento de dados, treinamento e avaliação dos modelos. Os algoritmos avaliados incluem *Decision Tree* (DT), *Gradient Boosting* (GB), *LightGBM* (LGBM), *Linear Support Vector* (LSV), *Logistic Regression* (LR), *Naive Bayes Multinomial* (NBM) e *Random Forest* (RF), todos implementados com a biblioteca *Sklearn* utilizando seus parâmetros padrão. Os modelos foram combinados à técnica de PLN *Term Frequency–Inverse Document Frequency* (TF-IDF) para representar textos de forma vetorial.

2. Fundamentação Teórica

De acordo com [Géron 2021], o ML é a ciência de programar computadores para que possam aprender com dados. Os modelos de ML são construídos a partir da observação dos dados de treinamento e utilizado para realizar novas previsões [Russell and Norvig 2022]. As abordagens incluem aprendizado supervisionado, aprendizado não supervisionado, semi-supervisionado e aprendizado por reforço [Kalinowski et al. 2023]. No aprendizado supervisionado, abordagem utilizada neste estudo, os algoritmos aprendem a partir de exemplos rotulados [Russell and Norvig 2022]. Segundo [Faceli et al. 2023], essa técnica pode ser subdividida em problemas de regressão, quando as saídas são numéricas, e de classificação, quando as saídas são categóricas.

No contexto da classificação, destacam-se três tipos: binária, com duas classes possíveis; multiclasse, com mais de duas classes, mas apenas uma por instância; e multirrótulo, que permite múltiplas classes para uma mesma instância [Vajjala et al. 2020, Faceli et al. 2023]. Esta última é especialmente útil no contexto das NFS-e, pois é comum que uma mesma nota fiscal descreva múltiplos serviços. Por exemplo, descrições como "consulta médica e exames laboratoriais" englobam múltiplas categorias de serviços, evidenciando a natureza multirrótulo do problema e, consequentemente, a necessidade da aplicação de técnicas apropriadas de classificação multirrótulos.

Devido à natureza textual das descrições contidas nas NFS-e, a aplicação de técnicas de Processamento de Linguagem Natural (PLN) são essenciais para transformação dessas informações em representações estruturadas, as quais viabilizam a aplicação de algoritmos de classificação capazes de identificar padrões e categorizar os serviços prestados. Isso porque o PLN é uma área da IA dedicada ao processamento computacional da linguagem humana [Caseli and Nunes 2024]. Entre as técnicas de PLN, destacam-se métodos de representação de texto como *One-Hot Encoding, Bag of Words, Bag of N-Grams*. Entretanto, elas não consideram a relevância das palavras [Vajjala et al. 2020].

Já o método TF-IDF (*Term Frequency–Inverse Document Frequency*) emprega técnicas estatísticas para quantificar a relevância de uma palavra para um documento ou coleção de documentos [Liu et al. 2018]. Segundo [Vajjala et al. 2020], essa abordagem combina duas medidas: a Frequência do Termo (TF), que avalia a frequência que um termo aparece em um documento específico, e a Frequência Inversa de Documentos (IDF), que indica o quão raro é um termo no *corpus*, assumindo que termos comuns tem menor poder para identificar um documento. O valor final de TF-IDF resulta do produto entre essas duas medidas, representada por [Liu et al. 2018] pela seguinte equação:

$$tfidf_{i,j} = \frac{tf_{i,j} \times idf_i}{\sqrt{\sum_{t_i \in d_j} [tf_{i,j} \times idf_i]^2}} \tag{1}$$

Apesar de amplamente utilizado, o TF-IDF apresenta limitações importantes. Por tratar os termos de maneira independente e ignorar a ordem das palavras, a técnica não captura as relações semânticas e contextuais. Além disso, ao gerar vetores esparsos e de alta dimensionalidade, o modelo pode resultar em maior custo computacional e ser mais suscetível a ruídos. Essas limitações evidenciam a importância de investigar, em estudos futuros, abordagens baseadas em *word embeddings*, que representam palavras de forma densa e contextualizada [Vajjala et al. 2020].

3. Trabalhos Relacionados

Diante da crescente demanda por maior controle fiscal e combate à evasão, pesquisas recentes têm explorado o uso de Inteligência Artificial (IA) na análise de documentos fiscais. Estudos anteriores demonstraram o potencial de algoritmos de ML, destacando o bom desempenho de classificadores como *Support Vector Machines* (SVM) e *Random Forest* (RF) em [Gomes 2023, Dias and Becker 2017, Soares and Cunha 2020], para classificar produtos, fraudes e inadimplências fiscais. No entanto, esses estudos se concentram em problemas binários ou multiclasse, sem considerar a complexidade da classificação multirrótulo presente nas NFS-e.

Paralelamente, abordagens baseadas em redes neurais, como CNNs e LSTMs, têm sido empregadas com bons resultados na identificação de padrões complexos em descrições de produtos, conforme demonstrado por [De Araujo Neto 2021]. Além disso, sistemas baseados em regras explícitas, como em [Lins Neto 2021], foram utilizados com sucesso em domínios específicos, embora enfrentem limitações na adaptação a variações nos dados, exigindo manutenções frequentes. A presente pesquisa propõe a aplicação de técnicas de IA na classificação de serviços descritos em NFS-e, com o objetivo de superar

essas limitações, promovendo maior adaptabilidade, precisão na detecção de irregularidades e redução da dependência de regras rígidas.

4. Metodologia

Os modelos foram desenvolvidos a partir do pipeline ilustrado na Figura 1, adaptado de [Vajjala et al. 2020]. Na aquisição de dados, foram coletadas aproximadamente 6,56 milhões de NFS-e do setor de saúde, emitidas entre 2022 e 2024. Em seguida, na etapa de rotulação de dados, cerca de 5,69 milhões foram rotuladas manualmente com base na análise textual do campo descritivo do serviço, resultando em 22 classes: 21 correspondentes aos códigos de serviços do setor de saúde e uma classe adicional foi criada para notas com descrições genéricas, nas quais não é possível identificar um serviço específico.

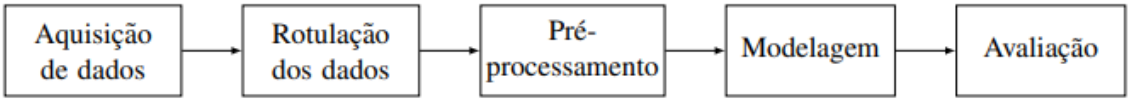


Figura 1. *Pipeline* do processo de construção dos modelos preditivos.

Na etapa de pré-processamento, foram aplicados quatro procedimentos fundamentais: normalização textual, com a conversão para letras minúsculas e remoção de acentuação; eliminação de caracteres e termos irrelevantes, como números, símbolos, letras isoladas e stopwords; separação de palavras concatenadas; e o balanceamento da base de dados, com o intuito de mitigar o desequilíbrio entre as classes. Para o balanceamento, foi adotada uma estratégia híbrida. Inicialmente, realizou-se a subamostragem aleatória sem reposição das classes majoritárias, utilizando a função **random.choice** da biblioteca Numpy. Em seguida, aplicou-se superamostragem com reposição das classes minoritárias por meio da função **utils.resample** da biblioteca Sklearn. Como resultado desse processo, a base balanceada passou a conter aproximadamente 2,1 milhões de amostras de NFS-e.

Na modelagem, a representação vetorial dos textos foi realizada por meio da técnica TF-IDF, utilizando apenas *uni-gramas*, conforme o padrão do método adotado. Essa abordagem implica que cada termo do vocabulário é tratado de forma isolada, desconsiderando combinações sequenciais de palavras. Além disso, foram mantidos os 10.000 termos mais relevantes, com intuito de equilibrar a qualidade da representação textual e a eficiência computacional. A construção dos modelos preditivos foram conduzidas por meio da técnica de validação cruzada *K-Fold*, conforme implementação disponível na biblioteca *Scikit-learn*, com cinco divisões (*n_splits=5*). Adicionalmente, foi empregado o embaralhamento dos dados (*shuffle=True*) e controle de aleatóriedade (*random_state=42*), de modo a assegurar a reprodutibilidade dos experimentos e a consistência dos resultados obtidos na comparação entre os diferentes algoritmos avaliados.

Por fim, na etapa de avaliação, o desempenho dos modelos foi analisado por meio das métricas de acurácia, precisão, *recall*, *F1-score* e *Hamming Loss* (HL), conforme proposto em [Faceli et al. 2023]. Com o objetivo de avaliar a eficiência computacional, também foram considerados o tempo médio de treinamento (TMT) e o tempo médio de predição (TMP), seguindo a abordagem descrita em [Gomes 2023]. Os resultados obtidos são apresentados e discutidos na seção seguinte.

5. Resultados

Nesta seção, são apresentados os resultados da avaliação comparativa dos modelos treinados. A Figura 2 indica que o *Random Forest* obteve a maior acurácia, com 99,3%. Em seguida, destacam-se os modelos *Decision Tree* (99,2%), *Linear Support Vector* (98,2%) e *LightGBM* (98%). No entanto, devido à natureza multirrótulodo problema, a acurácia isoladamente é insuficiente para uma avaliação conclusiva do desempenho. Por essa razão, é necessário recorrer a métricas complementares para uma análise mais abrangente.

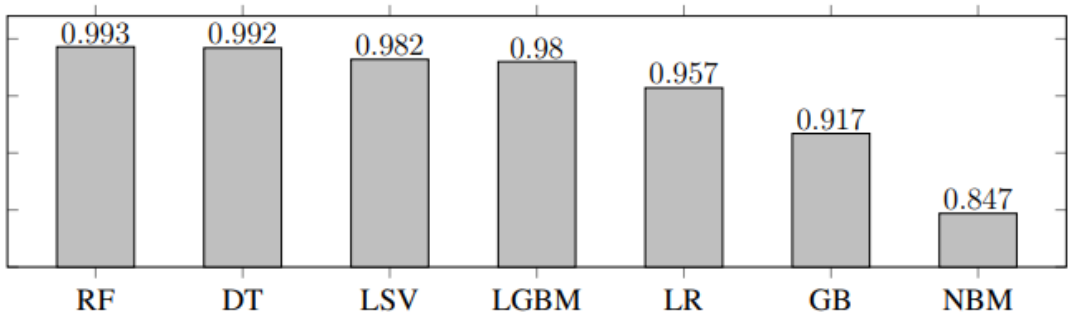


Figura 2. Análise Comparativa da Acurácia dos Modelos Treinados

A análise das demais métricas confirma o desempenho superior do modelo *Random Forest*, evidenciada por seus altos valores de precisão (0,9978), *recall* (0,9960) e *F1-score* (0,9969), além do menor índice de *Hamming Loss* (0,00039). Os resultados indicam elevada capacidade preditiva e robustez na classificação multirrótulo. Entretanto, esse desempenho vem acompanhado de um TMT significativamente elevado (01:47:36) e pelo maior TMP (00:01:29). Isso pode representar um fator limitante em aplicações que exigem eficiência computacional ou respostas em tempo reduzido. De maneira semelhante, o modelo *Decision Tree* também apresentou métricas elevadas, com *F1-score* de 0,9969 e uma *Hamming Loss* igualmente baixa (0,00045). Contudo, apesar de um TMT também elevado (01:45:05), o modelo apresentou baixo TMP (00:00:07).

Tabela 1. Desempenho dos Modelos

Modelo	Precisão	Recall	F1-Score	HL	TMT	TMP
Random Forest	0,9978	0,9960	0,9969	0,00039	01:47:36	00:01:29
Decision Tree	0,9971	0,9958	0,9965	0,00045	01:45:05	00:00:07
Linear Support Vector	0,9948	0,9914	0,9931	0,00093	00:10:28	00:00:03
LightGBM	0,9960	0,9901	0,9930	0,00101	00:15:06	00:00:12
Logistic Regression	0,9893	0,9758	0,9825	0,00220	00:01:09	00:00:03
Gradient Boosting	0,9917	0,9553	0,9719	0,00400	03:46:38	00:00:10
Naive Bayes Multinomial	0,9168	0,9490	0,9296	0,00921	00:00:05	00:00:04

Por outro lado, o modelo *Linear Support Vector* apresentou um desempenho competitivo, atingindo um *F1-score* de 0,9931 e *Hamming Loss* de apenas 0,00093. Destacou-se ainda, por apresentar os menores tempos tanto de treinamento (00:10:28) quanto de predição (00:00:03) entre os três modelos com melhor desempenho geral. De forma semelhante, o modelo *LightGBM* o modelo conciliou alta assertividade com eficiência computacional, alcançando um *F1-score* de 0,9930 e *Hamming Loss* de 0,00101, com TMT

em apenas 00:15:06 e TMP de 00:00:12. Essa combinação de alta performance preditiva e baixos custos computacionais torna ambos os modelos especialmente adequados para aplicações em ambientes caracterizados por restrições de recursos computacionais.

O *Naive Bayes Multinomial*, embora extremamente eficiente em termos de tempo de treinamento (00:00:05), apresentou limitações significativas em desempenho, com *F1-score* de 0,9296. O modelo obteve o pior desempenho geral, o que pode torná-lo inadequado ao contexto proposto. Já o modelo *Gradient Boosting* obteve desempenho considerado satisfatório, com *F1-score* de 0,9719. No entanto, apresentou o maior TMT, totalizando 03:46:38, o que pode comprometer sua viabilidade em cenários com restrições de tempo e recursos computacionais.

6. Conclusões e trabalhos futuros

Este estudo demonstrou a viabilidade da classificação multirrótulo de serviços em NFS-e do setor de saúde por meio de algoritmos de aprendizado de máquina. Os resultados indicaram que modelos baseados em árvores, como *Random Forest* e *Decision Tree*, alcançaram altos índices de desempenho, embora com maior custo computacional. Todavia, ao considerar a relação entre desempenho preditivo e eficiência computacional, os modelos *Linear Support Vector* e *LightGBM* destacaram-se como alternativa eficiente, combinando boa qualidade preditiva e baixo tempo de execução. Este experimento integra uma pesquisa mais ampla, cujo objetivo é desenvolver uma solução para automatizar a identificação dos serviços descritos na Prefeitura do Recife, com a finalidade de apoiar os processos de auditoria, além de automatizar alertas para os contribuintes sobre possíveis inconsistências.

Apesar dos resultados promissores, o estudo apresenta limitações relevantes: restringe-se ao setor de saúde e a NFS-e da Prefeitura do Recife, o que limita sua generalização a outros contextos econômicos e regionais. O processamento textual adotado pode não capturar nuances semânticas em descrições curtas ou ambíguas, e os modelos utilizados, baseados em algoritmos clássicos, não exploram técnicas mais avançadas, como redes neurais profundas ou *Large Language Models* (LLMs), que poderiam aprimorar o desempenho. Como direções para trabalhos futuros, os pesquisadores planejam expandir o escopo do estudo, incluindo serviços de outros setores, visto que o experimento atual está restrito ao setor de saúde. Além disso, há a intenção de explorar modelos baseados em *embeddings* e redes neurais profundas, bem como modelos de LLMs, visando aprimorar a generalização e a compreensão semântica dos textos.

Agradecimentos. Agradecemos à Prefeitura do Recife pelo apoio institucional e pelo comprometimento com a inovação em serviços públicos. Destacamos, em especial, a Secretaria de Finanças, pela disponibilização dos dados e pelo suporte técnico essencial ao desenvolvimento deste estudo.

Referências

ABRASF (2008). NFS-e: Modelo Conceitual. https://abrasf.org.br/biblioteca/arquivos-publicos/nfs-e/versao-1-00. Acesso em: 10 out. 2024.

Caseli, H. M. and Nunes, M. G. V. (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 edition.

De Araujo Neto, A. M. (2021). O uso de processamento de linguagem natural para classificação de produtos no contexto de notas fiscais eletrônicas. 25 p. Trabalho de Conclusão de Curso - Universidade Estadual da Paraíba, Campina Grande.

De Macedo, C. and Diniz Filho, J. W. d. F. (2019). Sonegação Fiscal: Uma análise dos seus Efeitos na Economia Brasileira. *Revista de Auditoria Governança e Contabilidade*, 7(31). Acesso em: 16 out. 2024.

Dias, M. and Becker, K. (2017). Identificação de Candidatos à Fiscalização por Evasão do Tributo ISS. *Anais do 5º Symposium on Knowledge Discovery, Mining and Learning*.

Dos Anjos, P. G. and Pinheiro, M. T. S. (2024). A implementação da inteligência artificial (ia) na fiscalização tributária: inovações disruptivas para eficiência na arrecadação do iptu. *Revista Tributária e de Finanças Públicas*, 159. Acesso em: 30 nov. 2024.

Faceli, K. et al. (2023). *Inteligência Artificial: uma abordagem de aprendizado de máquina*. LTC, Rio de Janeiro, 2 edition.

Gomes, W. F. (2023). Análise exploratória e experimental de aplicações de inteligência artificial para classificação de descrições incongruentes em compras na área de saúde pública. Dissertação, Universidade Federal de Sergipe, São Cristóvão.

Géron, A. (2021). *Mãos à Obra Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow: Conceitos, Ferramentas e Técnicas Para a Construção de Sistemas Inteligentes*. Alta Books, Rio de Janeiro.

Kalinowski, M. et al. (2023). *Engenharia de Software para Ciência de Dados: Um guia de boas práticas com ênfase na construção de sistemas de machine learning*. Casa do Código, São Paulo.

Lins Neto, J. C. (2021). Audita-nfse: sistema auxiliar de auditoria em notas fiscais de serviços eletrônicas. Dissertação, Universidade Federal de Pernambuco, Recife.

Liu, C.-z., Sheng, Y.-x., Wei, Z.-q., and Yang, Y.-Q. (2018). Research of text classification based on improved tf-idf algorithm. In *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, pages 218–222.

Neto, H. D. A. and Martinez, A. L. (2016). Nota Fiscal de Serviços Eletrônica: Uma análise dos impactos na arrecadação em municípios brasileiros. *Revista de Contabilidade e Organizações*, 10(26):49–62.

Pinheiro, G. J. and Cunha, L. R. S. (2009). A importância da auditoria na detecção de fraudes. *Contabilidade Vista amp; Revista*, 14(1):31–48.

Russell, S. and Norvig, P. (2022). *Inteligência Artificial: uma abordagem moderna*. Grupo Editora Nacional, Rio de Janeiro, 4 edition.

Soares, G. and Cunha, R. (2020). Predição de Irregularidade Fiscal dos Contribuintes do Tributo ISS. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, pages 223–228, Porto Alegre, RS, Brasil. SBC.

Vajjala, S. et al. (2020). *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media, Sebastopol, CA, USA.

APÊNDICE B – Artigo: Nota Conforme: Sistema Integrado para Classificação Automatizada de Serviços em NFS-e com Machine Learning.

Nota Conforme: Sistema Integrado para Classificação Automatizada de Serviços em NFS-e com Machine Learning

Tarcísio Paraiso Farias[1], **Thiago Souto Mendes**[1], **Rafael Sena da Conceição** [2]

[1]Programa de Pós-Graduação em Engenharia de Sistemas e Produtos (PPGESP)
Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA)

tarcisioparaiso@gmail.com, thiagosouto@ifba.edu.br

[2]Secretaria de Finanças - Prefetura do Recife (PE)

rafaelsena@gmail.com

Abstract. *This article presents Nota Conforme, a system developed to support the Recife-PE City Hall in the inspection processes of Electronic Service Invoices (NFS-e). Through machine learning, the system automatically classifies the services described in the NFS-e documents issued within the healthcare sector. The tool offers integration with the city hall's internal systems via API, aiming to assist auditors in identifying possible tax fraud, as well as enabling the automation of tasks such as sending emails and communicating with the NFS-e issuer, in order to alert taxpayers about potential inconsistencies.*

Resumo. *Este artigo apresenta o Nota Conforme, um sistema desenvolvido para apoiar a Prefeitura do Recife-PE nos processos de fiscalização em Notas Fiscais de Serviços Eletrônicas (NFS-e). Por meio de aprendizado de máquina, o sistema classifica automaticamente os serviços descritos nas NFS-e do setor de saúde. A ferramenta oferece integração com sistemas internos da prefeitura por meio de API, para auxiliar os auditores na identificação de possíveis fraudes fiscais. Além disso, possibilita a automatização de tarefas como o envio de e-mails e a comunicação com o próprio emissor de NFS-e, com o objetivo de notificar os contribuintes sobre eventuais inconsistências identificadas.*

1. Introdução

A Nota Fiscal de Serviços Eletrônica (NFS-e) é um documento digital que registra operações de prestação de serviços [ABRASF 2008], contribuindo para a eficiência da administração tributária [Neto and Martinez 2016]. Contudo, práticas fraudulentas visam reduzir tributos [Dias and Becker 2017], configurando sonegação fiscal [BRASIL 1965] e prejudicando investimentos sociais [De Macedo and Diniz Filho 2019]. Assim, o monitoramento eficaz é essencial para mitigar esse problema [Pinheiro and Cunha 2009].

Entretanto, dada a inviabilidade da análise humana diante da complexidade e do volume das bases de dados municipais, a implementação de ferramentas de Inteligência Artificial (IA) pode potencializar a extração de informações relevantes e aprimorar significativamente os processos de fiscalização [Dos Anjos and Pinheiro 2024]. Diante disso, este estudo tem como objetivo apresentar o sistema Nota Conforme[1], desenvolvido para atender à demanda da Prefeitura do Recife-PE por uma solução automatizada

[1]Vídeo de demonstração: https://youtu.be/OLaRtmaSyXY e repositório com o código fonte: https://github.com/tpfarias/notaconforme

na classificação dos serviços descritos nas NFS-e do setor de saúde, o segmento foi escolhido inicialmente devido à sua relevância apontada por auditores, especialmente em relação ao volume de notas e à ocorrência de inconsistências.

Seu principal objetivo é apoiar a fiscalização, auxiliando na identificação de divergências entre os serviços declarados e as alíquotas aplicadas. Além disso, o Nota Conforme permite a integração com sistemas internos de monitoramento, viabilizando o envio de alertas e relatórios de inconsistências para os contribuintes. Dessa forma, a proposta visa contribuir para a modernização da gestão tributária municipal, ao promover maior eficiência, transparência e assertividade nas ações de fiscalização.

2. Trabalhos Relacionados

Estudos recentes investigaram o uso de técnicas computacionais na classificação de dados fiscais. [Gomes 2023] propôs um sistema de *machine learning* (ML) para classificar Órteses, Próteses e Materiais Especiais com base na descrição de produtos em NF-e, por meio de interface para entrada textual e exibição de resultados. Já [Lins Neto 2021] desenvolveu um conjunto de regras para identificar NFS-e suspeitas de fraude na Construção Civil, implementadas em sistema que processa arquivos CSV e gera saídas classificadas.

Outros trabalhos focaram em modelos preditivos para detectar fraudes fiscais [De Araujo Neto 2021, Dias and Becker 2017, Soares and Cunha 2020], sem propor sistema para aplicação prática. Diferente desses estudos, este avança com duas contribuições principais: (i) abordagem centrada na classificação multirrótulo de códigos de serviço, oferecendo maior flexibilidade e durabilidade na análise fiscal; e (ii) desenvolvimento de API de predição que facilita a integração da solução com sistemas diversos.

3. O Sistema Nota Conforme

3.1. Modelo Preditivo

O modelo preditivo foi escolhido com base na avaliação comparativa de sete algoritmos de ML, conforme apresentado na Tabela 1. A base de dados utilizada no experimento compreende aproximadamente 5,69 milhões de NFS-e referentes a serviços da área de saúde, emitidas entre 2022 e 2024 na Prefeitura do Recife-PE. As NFS-e foram rotuladas manualmente com base na descrição, resultando em 21 serviços do setor. Devido à presença de descrições genéricas que impedem a identificação de um serviço, criou-se uma 22ª classe para tais instâncias.

Tabela 1. Desempenho dos Modelos Preditivos

Modelo	Precisão	Recall	F1-Score	HL	TMT	TMP
Random Forest	0,9978	0,9960	0,9969	0,00039	01:47:36	00:01:29
Decision Tree	0,9971	0,9958	0,9965	0,00045	01:45:05	00:00:07
Linear SVM	0,9948	0,9914	0,9931	0,00093	00:10:28	00:00:03
LightGBM	0,9960	0,9901	0,9930	0,00101	00:15:06	00:00:12
Logistic Regression	0,9893	0,9758	0,9825	0,00220	00:01:09	00:00:03
Gradient Boosting	0,9917	0,9553	0,9719	0,00400	03:46:38	00:00:10
Naive Bayes Multinomial	0,9168	0,9490	0,9296	0,00921	00:00:05	00:00:04

TMT: Tempo Médio de Treinamento. **TMP**: Tempo Médio de Predição. **HL**: *Hamming Loss.*

A base original apresentava forte desequilíbrio, com classes concentrando até 25% das amostras, enquanto outras tinham menos de 0,1%. Com o balanceamento, a base passou a ter 2,1 milhões de amostras com distribuição mais uniforme, cujas proporções variam de 4,6% a 11,35%. Os experimentos foram realizados por meio técnica de validação cruzada *K-Fold*, com 5 divisões (*n_splits=5*), implementada na biblioteca *Sklearn*. Para assegurar a reprodutibilidade dos resultados, adotou-se embaralhamento prévio dos dados (*shuffle=True*) e definição explícita do estado aleatório (*random_state* = 42).

O modelo *Random Forest* foi selecionado por apresentar o melhor desempenho geral entre os algoritmos avaliados, com os maiores valores de precisão (0,9978), *recall* (0,9960) e *F1-score* (0,9969). Além disso, obteve o menor valor de *Hamming Loss* (0,00039), evidenciando baixa taxa de erro na classificação multirrótulo.

3.2. Arquitetura do Sistema

O Diagrama de Contexto fornece uma visão geral das integrações da solução [C4 Model 2025], com destaque para o *pipeline* de dados que alimenta o *Data Warehouse* da NFS-e, apoiando a atuação fiscal. O sistema também pode se integrar ao emissor de NFS-e, permitindo a detecção de inconsistências durante a emissão, e ao módulo de envio de e-mails, automatizando a comunicação de inconformidades aos contribuintes.

Figura 1. Diagrama de contexto do Nota Conforme

O Diagrama de Contêiner detalha a arquitetura do sistema ao evidenciar a distribuição de responsabilidades entre seus principais componentes [C4 Model 2025]. No Nota Conforme, destacam-se três contêineres: a API, desenvolvida em *Python* com *FastAPI*, que integra sistemas externos, acessa o banco de dados e aciona o modelo preditivo; o Banco de Dados, em PostgreSQL 16.2, exclusivo à API, que armazena dados dos usuários; e o Modelo Preditivo, em *Python*, baseado em *Random Forest* com TF-IDF (*Term Frequency-Inverse Document Frequency*), que realiza a classificação dos códigos de serviço nas NFS-e a partir dos dados recebidos da API.

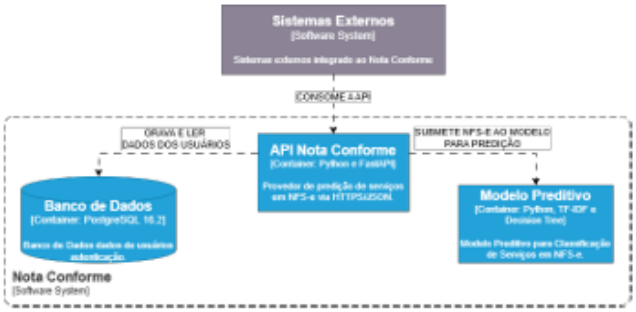


Figura 2. Diagrama de Contêiner do Nota Conforme

Já o Diagrama de Componentes detalha a estrutura interna do sistema, evidenciando os elementos e responsabilidades de cada contêiner [C4 Model 2025]. Na API do Nota Conforme, destacam-se três componentes principais: o CRUD[2] de Usuários, responsável pelas operações básicas de manipulação de dados; o Componente de Autenticação, que valida credenciais e emite tokens JWT; e o Componente de Predição, que processa NFS-e para identificação dos serviços via modelo preditivo. Esse diagrama é fundamental para compreender a organização e as tecnologias da aplicação.

Figura 3. Diagrama de componentes do Nota Conforme

3.3. Funcionalidades e Demonstração

A Figura 4 apresenta os dois recursos da API do Nota Conforme: **Usuarios** e **Predicoes**. O recurso **Usuarios** contempla funcionalidades de cadastro, alteração, exclusão, recuperação de dados e autenticação. Já o recurso **Predicoes**, recebe conjuntos de NFS-e para classificação dos serviços prestados com base na descrição das notas. Além disso, a figura também detalha os *endpoints* e métodos HTTP associados a cada operação.

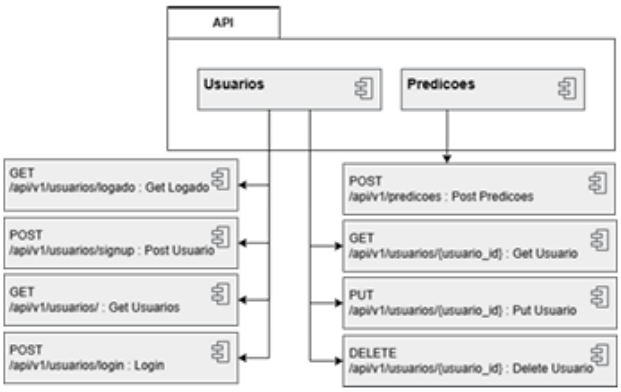


Figura 4. Recursos da API Nota Conforme

Todos os *endpoints* são protegidos e necessitam de autenticação para acesso aos recursos. Através do *endpoint* **login** é realizada a autenticação com usuário e senha por meio do método POST para obtenção do *token* JWT. O *token* será útil para acessar os demais *endpoints*, como o **predicoes**, que além do *token*, deverá enviar os dados das NFS-e que serão classificadas na requisição. Dessa forma, a API responderá com as predições realizadas pelo modelo.

[2]CRUD - Acrônimo para as quatro operações básicas de manipulação de dados: *Create* (Criar), *Read* (Ler), *Update* (Atualizar) e Delete (Excluir).

Para realizar predições de serviços em NFS-e usando o Nota Conforme, é necessário incluir ao menos o campo **discriminacao** na requisição. Os demais campos são opcionais e podem ser usados conforme a necessidade de cada organização. Conforme apresentado na Figura 5, a resposta da API incluirá o campo **servicos**, com os códigos de serviços identificados na descrição da nota, além dos demais campos enviados, exceto **discriminacao**, que é omitido para reduzir o volume de dados. Isso garante uma solução flexível e compatível com diferentes sistemas.

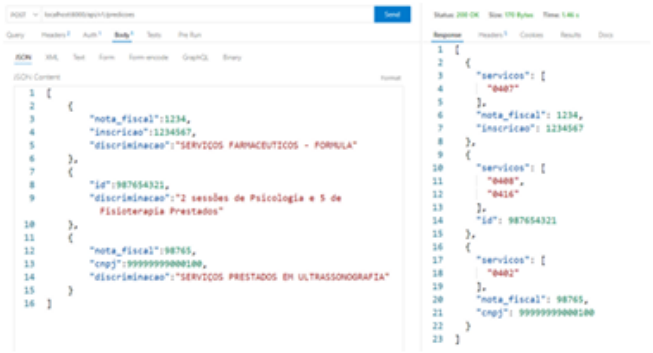


Figura 5. Exemplo de requisição e resposta do Nota Conforme

O sistema Nota Conforme foi avaliado em colaboração com auditores da Prefeitura do Recife–PE, por meio da análise de 500 NFS-e selecionadas aleatoriamente. Dois auditores participaram do processo de avaliação. Eles foram responsáveis tanto pela seleção das notas fiscais quanto pela análise manual de desempenho, realizada com base na comparação entre as predições do modelo e a classificação atribuída por eles. A solução apresentou desempenho satisfatório, alcançando um *F1-Score* de 0,945.

Adicionalmente, o processamento da amostra foi concluído em apenas 3,45 segundos, demonstrando boa eficiência computacional. Esse desempenho foi obtido em um computador com processador Intel Core i7-1260P (12 núcleos, 16 threads), 16 GB de RAM e GPU integrada Intel Iris Xe. Embora os resultados sejam promissores para aplicações em tempo real com baixa latência, em cenários com grandes volumes de dados ou ambientes de produção, análises complementares de uso de CPU, memória e paralelização são recomendadas para avaliar sua viabilidade em larga escala.

4. Considerações Finais

O sistema Nota Conforme demonstrou ser uma solução eficaz no apoio aos auditores fiscais, ao facilitar a detecção de inconformidades por meio da classificação automática dos serviços descritos nas NFS-e. A abordagem baseada na identificação dos serviços revelou-se robusta e duradoura, uma vez que não depende de regras fiscais específicas ou temporais, diferentemente de métodos anteriores focados apenas na detecção de fraudes.

A escolha do modelo preditivo foi fundamentada em testes comparativos entre diferentes algoritmos de aprendizado de máquina, considerando tanto a sua assertividade quanto a eficiência computacional. Além disso, o sistema se mostrou promissor ao oferecer integração com diversas ferramentas internas por meio de uma API, ampliando sua aplicabilidade prática.

Como direções para trabalhos futuros, propõe-se ampliar o escopo para incluir serviços de outros setores. Também está prevista o desenvolvimento de uma plataforma para gerir as inconsistências encontradas, bem como incorporar novos atributos da NFS-e que possibilitem o cálculo dos valores evadidos e recuperados, reforçando o suporte à fiscalização. Por fim, planeja-se avaliar abordagens baseadas em *embeddings* para melhor capturar relações semânticas complexas e aprimorar o desempenho do modelo.

Agradecimentos. Este estudo contou com o apoio da Prefeitura do Recife-PE, via Secretaria de Finanças, que forneceu dados e equipe técnica para o projeto.

Referências

ABRASF (2008). NFS-e: Modelo Conceitual. Versão 1. https://abrasf.org.br/biblioteca/arquivos-publicos/nfs-e/versao-1-00. Acesso em: 10 out. 2024.

BRASIL (1965). Lei nº 4.729, de 14 de julho de 1965. define o crime de sonegação fiscal e dá outras providências. https://www.planalto.gov.br/ccivil_03/leis/1950-1969/l4729.htm. Acesso em: 20 jul. 2024.

C4 Model (2025). C4 model for visualising software architecture. https://c4model.com/. Acesso em: 12 abr. 2025.

De Araujo Neto, A. M. (2021). O uso de processamento de linguagem natural para classificação de produtos no contexto de notas fiscais eletrônicas. Trabalho de Conclusão de Curso de Graduação - Universidade Estadual da Paraíba, Campina Grande.

De Macedo, C. and Diniz Filho, J. W. d. F. (2019). Sonegação Fiscal: Uma análise dos seus Efeitos na Economia Brasileira. *Revista de Auditoria Governança e Contabilidade*, 7(31). Acesso em: 16 out. 2024.

Dias, M. and Becker, K. (2017). Identificação de Candidatos à Fiscalização por Evasão do Tributo ISS. *Anais do 5º Symposium on Knowledge Discovery, Mining and Learning*.

Dos Anjos, P. G. and Pinheiro, M. T. S. (2024). A implementação da inteligência artificial (ia) na fiscalização tributária: inovações disruptivas para eficiência na arrecadação do iptu. *Revista Tributária e de Finanças Públicas*, 159. Acesso em: 30 nov. 2024.

Gomes, W. F. (2023). Análise exploratória e experimental de aplicações de inteligência artificial para classificação de descrições incongruentes em compras na área de saúde pública. Dissertação, Universidade Federal de Sergipe, São Cristóvão.

Lins Neto, J. C. (2021). Audita-nfse: sistema auxiliar de auditoria em notas fiscais de serviços eletrônicas. Dissertação, Universidade Federal de Pernambuco, Recife.

Neto, H. D. A. and Martinez, A. L. (2016). Nota Fiscal de Serviços Eletrônica: Uma análise dos impactos na arrecadação em municípios brasileiros. *Revista de Contabilidade e Organizações*, 10(26):49–62.

Pinheiro, G. J. and Cunha, L. R. S. (2009). A importância da auditoria na detecção de fraudes. *Contabilidade Vista amp; Revista*, 14(1):31–48.

Soares, G. and Cunha, R. (2020). Predição de Irregularidade Fiscal dos Contribuintes do Tributo ISS. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, pages 223–228, Porto Alegre, RS, Brasil. SBC.

APÊNDICE C – Plano de teste e validação do modelo em conjunto com auditores.

INTRODUÇÃO

Estudante Pesquisador:	Tarcisio Paraiso Farias
Orientador:	Dr. Thiago Mendes
Instituição de Ensino:	Instituto Federal da Bahia - IFBA
Programa:	Programa de Pós-Graduação em Engenharia de Sistemas e Produtos
Entidade Parceira:	Secretaria de Finanças da Prefeitura Municipal do Recife–PE.
Objetivo:	Validar a capacidade dos modelos em classificar serviços por meio da observação da discriminação da nota fiscal. Comparar as classificações automáticas do modelo com a análise humana realizada pelos auditores.
Hipótese:	Os modelos são capazes de identificar corretamente os serviços descritos na discriminação da nota fiscal de serviço (NFS-e)?
Meta:	Garantir que o modelo escolhido seja eficaz e útil para a aplicação prática na auditoria fiscal.
Quantidade de participantes:	3 participantes
Colaboradores da Entidade Parceira:	Aline Assis, Leonardo Silva e Rafael Sena
Perfis dos participantes:	Auditores do Tesouro do Município do Recife–PE.
Métricas de avaliação:	Matriz de Confusão, Acurácia, Precisão, Recall, F1-Score e Especificidade.

PROCEDIMENTOS

Disponibilização da base de NFS-e

Serão extraídas da base de dados todas as notas emitidas entre os dias 01/04/2024 e 20/09/2024. As notas serão de serviços do setor de Saúde, representadas pelos códigos que variam entre 0401 e 0421. Em seguida, a base será exportada e enviada para os auditores em formata suportado por planilha eletrônica, com os seguintes campos:

CAMPO	DESCRIÇÃO
AFNFSESEQU	Sequencial da tabela fato de NFS-e no *Data Warehouse*
INSCRICAO	Número da Inscrição do contribuinte que emitiu a NFS-e
NOTA_FISCAL	Número da NFS-e
DATA_EMISSAO	Data de emissão da NFS-e
CNAE	Número do CNAE
CNAE_DESCRICAO	Descrição do CNAE
SERVICO_COD	Código do serviço municipal
SERVICO_DESC	Descrição do serviço municipal
DISCRIMINACAO	Discriminação da NFS-e

Seleção das NFS-e

Os auditores selecionarão a quantidade mínima de 500 notas, que deverão representar diferentes tipos de serviços do setor de Saúde. Entretanto, os auditores estarão livres para selecionar um volume maior de notas, se julgarem viável para análise manual da correspondência dos códigos de serviços. Independentemente da quantidade de notas selecionadas, recomenda-se aos auditores que escolham 50% das notas emitidas com o tipo de serviço correspondente à discriminação do serviço, e os outros 50% das notas inconsistentes, nas quais o tipo de serviço selecionado para fins tributários é divergente da discriminação do serviço.

Os auditores disponibilizarão a planilha contendo todas as notas selecionadas. Para garantir a imparcialidade dos testes, o arquivo deve conter somente as colunas enviadas na planilha original, sem nenhuma informação que possa identificar a análise manual dos auditores.

Submissão das NFS-e ao modelo para classificação

Após o recebimento da planilha, as NFS-e serão submetidas ao modelo de classificação. Com os resultados das classificações geradas pelo modelo, será acrescentado um novo campo

à planilha original contendo os códigos dos serviços atribuídos pelo modelo. Em seguida, a planilha será devolvida para os auditores.

Avaliação de assertividade do modelo

Após o recebimento da nova planilha, os auditores deverão comparar o desempenho do modelo com suas análises manuais. Durante o processo de avaliação, eles acrescentarão dois novos campos à planilha. O primeiro indicará se o modelo classificou corretamente os serviços descritos na NFS-e. O segundo campo será utilizado para registrar o correto quando o modelo falhar na classificação, além de outras observações que os auditores julgarem necessárias. Em seguida, a planilha com a análise dos auditores deve ser devolvida.

Devolução dos resultados

Com a análise dos auditores, as métricas de avaliação serão calculadas e os resultados serão apresentados. Com base nos resultados da validação, a Secretaria de Finanças poderá decidir sobre a viabilidade de implantação do modelo por meio da integração com Data Warehouse (DW) de NFS-e.

APÊNDICE D – Questionário para estudo de aceitação e viabilidade do Nota Conforme aplicados aos potenciais usuários.

ESTUDO DE VIABILIDADE DA APLICAÇÃO DO NOTA CONFORME

Este questionário faz parte de um **estudo de viabilidade** da utilização do sistema **Nota Conforme**, uma solução baseada em Inteligência Artificial voltada à **identificação automática dos serviços descritos** nas Notas Fiscais de Serviços Eletrônicas (**NFS-e**). O **objetivo** desta atividade é **coletar** suas **percepções** sobre o **potencial uso da ferramenta como apoio à fiscalização**, por meio do preenchimento das questões a seguir.

As respostas serão utilizadas **exclusivamente** para fins de **pesquisa** e **tratadas** de forma **confidencial**.

PERFIL DO PARTICIPANTE

1. Nome do participante:

2. Cargo atual:

3. Tempo de experiência com auditorias fiscais/tributárias (em anos):

4. Experiência com análise de NFS-e:

(1) Nenhuma experiência ou familiaridade	(2) Pouca experiência ou familiaridade	(3) Experiência ou familiaridade moderada	(4) Boa experiência ou familiaridade	(5) Muita experiência ou domínio do tema

AFIRMAÇÕES	1	2	3	4	5
Tenho experiência com análise do campo de discriminação de serviços nas NFS-e.					
Estou familiarizado com inconsistências comuns na descrição de serviços em NFS-e.					

5. Familiaridade com tecnologias aplicadas à fiscalização:

(1) Nenhuma experiência ou familiaridade	(2) Pouca experiência ou familiaridade	(3) Experiência ou familiaridade moderada	(4) Boa experiência ou familiaridade	(5) Muita experiência ou domínio do tema

AFIRMAÇÕES	1	2	3	4	5
Já utilizei ou tive contato com ferramentas tecnológicas de apoio à auditoria fiscal.					
Tenho facilidade em interpretar relatórios ou dashboards oriundos de sistemas de apoio à fiscalização.					
Estou familiarizado com o uso de soluções baseadas em inteligência artificial ou aprendizado de máquina no contexto da administração tributária.					

6. Percepção sobre inovação tecnológica no trabalho:

(1) Discordo Totalmente	(2) Discordo Parcialmente	(3) Neutro	(4) Concordo parcialmente	(5) Concordo Totalmente

	1	2	3	4	5
Acredito que tecnologias como IA e automação podem melhorar significativamente os processos de fiscalização tributária.					
Estou disposto a incorporar novas tecnologias que otimizem o trabalho de auditoria, desde que os resultados sejam confiáveis.					

QUESTIONÁRIO DE AVALIAÇÃO

7. Em relação à **utilidade** do Nota Conforme, marque a opção que melhor representa seu ponto de vista:

(1) Discordo Totalmente	(2) Discordo Parcialmente	(3) Neutro	(4) Concordo parcialmente	(5) Concordo Totalmente

	1	2	3	4	5
A identificação automática dos serviços descritos nas NFS-e tem potencial para aumentar minha produtividade como auditor.					
As informações que a solução se propõe a gerar poderão tornar o processo de fiscalização mais eficiente.					
A ferramenta pode ajudar a detectar inconsistências que eu poderia não identificar manualmente.					
Acredito que o uso da ferramenta contribui para a qualidade das análises realizadas na auditoria fiscal.					
A automatização de alertas diretamente no emissor de NFS-e contribui para a prevenção de erros durante a emissão.					
A automatização da comunicação de inconformidades por e-mail melhora a efetividade do processo de fiscalização.					

8. Em relação à **facilidade de uso** do Nota Conforme, marque a opção que melhor representa seu ponto de vista:

(1) Discordo Totalmente	(2) Discordo Parcialmente	(3) Neutro	(4) Concordo parcialmente	(5) Concordo Totalmente

	1	2	3	4	5
As informações geradas pela solução são apresentadas de forma clara e compreensível.					
Seria fácil compreender e utilizar as informações produzidas pela ferramenta.					
Os dados produzidos pela solução se integram bem às minhas ferramentas ou rotinas atuais.					
A adoção dessa ferramenta não sobrecarrega meu trabalho com tarefas adicionais.					

10. Em relação a um **possível uso futuro** do Nota Conforme, marque a opção que melhor representa seu ponto de vista:

(1) Discordo Totalmente	(2) Discordo Parcialmente	(3) Neutro	(4) Concordo parcialmente	(5) Concordo Totalmente

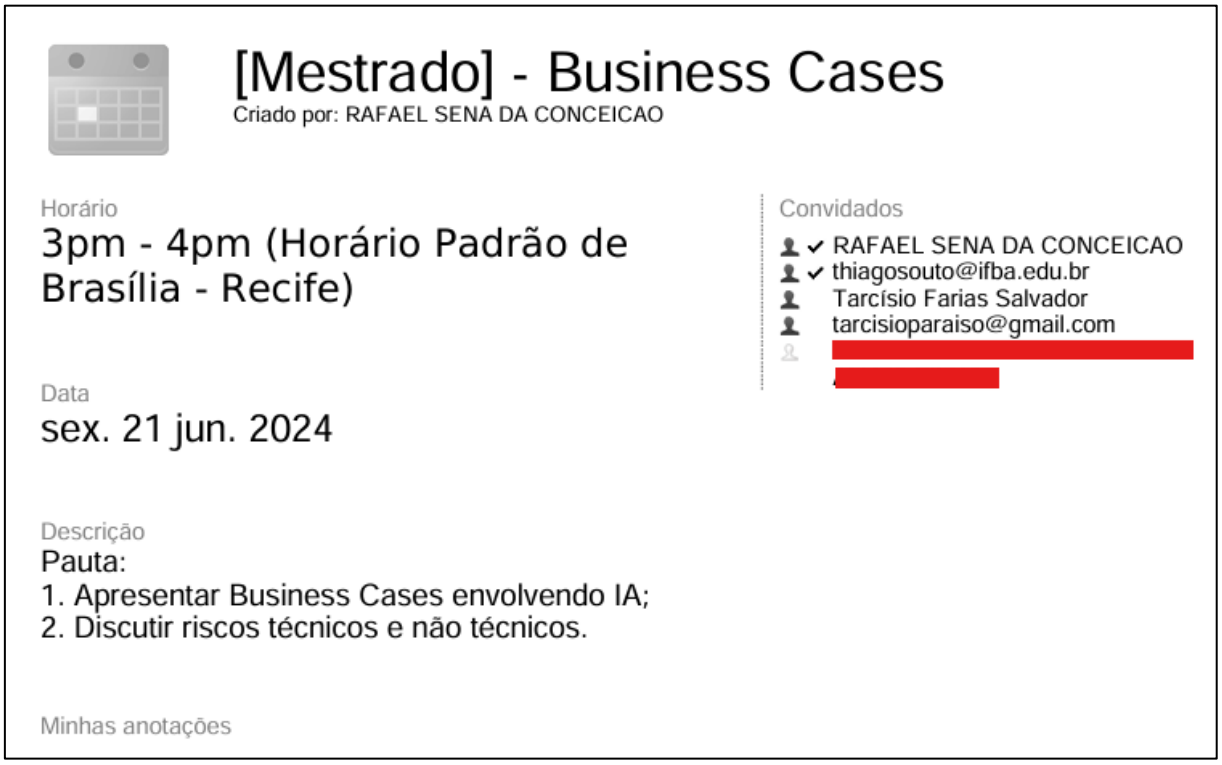
	1	2	3	4	5
Pretendo usar as informações geradas por essa solução sempre que estiverem disponíveis.					
Considero que os dados produzidos pela ferramenta podem ser incorporados à rotina de auditoria após sua homologação.					
Estou disposto a utilizar os resultados dessa ferramenta como base para parte da minha tomada de decisão.					
Prefiro realizar a fiscalização com apoio da identificação automática dos serviços do que com o processo manual atualmente adotado.					

11. De acordo com sua opinião, liste os aspectos **positivos** da utilização do Nota Conforme **(opcional)**:

12. De acordo com sua opinião, liste os aspectos **negativos** da utilização do Nota Conforme **(opcional)**:

13. Você possui alguma **sugestão para melhoria** do Nota Conforme? Em caso positivo, por favor, especifique-a **(opcional)**:

APÊNDICE E – Convites e pautas de reuniões realizadas ao longo do desenvolvimento do projeto.



[Mestrado] - Business Cases

Criado por: RAFAEL SENA DA CONCEICAO

Horário
3pm - 4pm (Horário Padrão de Brasília - Recife)

Convidados
- ✓ RAFAEL SENA DA CONCEICAO
- ✓ thiagosouto@ifba.edu.br
- Tarcísio Farias Salvador
- tarcisioparaiso@gmail.com
- [redacted]

Data
sex. 21 jun. 2024

Descrição
Pauta:
1. Apresentar Business Cases envolvendo IA;
2. Discutir riscos técnicos e não técnicos.

Minhas anotações

[Mestrado] - Saúde

Criado por: RAFAEL SENA DA CONCEICAO · Sua resposta: ✓Sim, eu vou

Horário
11am - 12pm (Horário Padrão de Brasília - Recife)

Convidados
- ✓ [redacted]
- ✓ RAFAEL SENA DA CONCEICAO
- ✓ Tarcísio Farias Salvador

Data
sex. 5 jul. 2024

Descrição
Pauta:
1. Apresentar as maiores dificuldades/complexidades da fiscalização da saúde;
2. Discutir possíveis soluções de IA que ajudem com a resolução dos problemas elencados no item 1;
3. O que ocorrer.

Minhas anotações

[Mestrado] - Construção civil

Criado por: RAFAEL SENA DA CONCEICAO · Sua resposta: ✔Sim, eu vou

Horário
10am - 11am (Horário Padrão de Brasília - Recife)

Convidados
- ✔ [redacted]
- ✔ [redacted]
- ✔ RAFAEL SENA DA CONCEICAO
- ✔ Tarcísio Farias Salvador

Data
sex. 12 jul. 2024

Descrição
HORÁRIO SUJEITO A RATIFICAÇÃO

Pauta:
1. Apresentar as maiores dificuldades/complexidades da fiscalização da construção civil;
2. Discutir possíveis soluções de IA que ajudem com a resolução dos problemas elencados no item 1;
3. O que ocorrer.

Minhas anotações

[Mestrado] - ML na saúde

Criado por: RAFAEL SENA DA CONCEICAO

Horário
11am - 12pm (Horário Padrão de Brasília - Recife)

Convidados
- ✔ [redacted]
- ✔ [redacted]
- ✔ RAFAEL SENA DA CONCEICAO
- [redacted]
- Tarcísio Farias Salvador
- [redacted]
- thiagosouto@ifba.edu.br

Data
qui. 5 set. 2024

Descrição
Pauta:
1.Apresentar resultados de negócio preliminares.
2. Apresentar a forma de implementação do modelo ML.

Minhas anotações

[Mestrado] - ML na saúde

Criado por: RAFAEL SENA DA CONCEICAO

Horário
11am - 12pm (Horário Padrão de Brasília - Recife)

Data
seg. 23 set. 2024

Convidados
[redacted]
RAFAEL SENA DA CONCEICAO
[redacted]
Tarcísio Farias Salvador
[redacted]
thiagosouto@ifba.edu.br

Descrição
Pauta:
1.Apresentar a 2a rodada de resultados de negócio preliminares;
2. O que ocorrer.

[Mestrado] - Apresentação preliminar dos resultados

Criado por: RAFAEL SENA DA CONCEICAO · Sua resposta: ✔Sim, eu vou

Horário
3pm - 4pm (Horário Padrão de Brasília - Bahia)

Data
qua. 8 jan. 2025

Convidados
✔ RAFAEL SENA DA CONCEICAO
✔ Tarcísio Farias Salvador

[MESTRADO] - Dúvidas sobre as análises

Criado por: RAFAEL SENA DA CONCEICAO

Horário
10am - 10:30am (Horário Padrão de Brasília - Recife)

Data
sex. 10 jan. 2025

Convidados
✔ [redacted]
✔ RAFAEL SENA DA CONCEICAO
Tarcísio Farias Salvador

[MESTRADO] - Dúvidas sobre a validação

Criado por: RAFAEL SENA DA CONCEICAO · Sua resposta: ✔Sim, eu vou

Horário
10:30am - 11:30am (Horário Padrão de Brasília - Recife)

Convidados
✔ RAFAEL SENA DA CONCEICAO
✔ Tarcísio Farias Salvador
[redacted]

Data
ter. 4 fev. 2025

[MESTRADO] - Apresentação dos Resultados Finais

Criado por: RAFAEL SENA DA CONCEICAO · Sua resposta: ✔Sim, eu vou

Horário
10:30am - 12pm (Horário Padrão de Brasília - Recife)

Data
qui. 27 fev. 2025

Descrição
Pauta:
1. Apresentar os resultados;
2. Apresentar o processo;
3. Sanar dúvidas;
4. Levantar as oportunidades.

Convidados
[redacted]
✔ RAFAEL SENA DA CONCEICAO
✔ Tarcísio Farias Salvador
[redacted]

[MESTRADO] - PASSAGEM DE CONHECIMENTO API

Criado por: RAFAEL SENA DA CONCEICAO

Horário
8am - 9am (Horário Padrão de Brasília - Recife)

Convidados
✔ RAFAEL SENA DA CONCEICAO
Tarcísio Farias Salvador

Data
sex. 21 mar. 2025

ANEXOS

ANEXO A – E-mail de Aceite do Artigo "Nota Conforme: Sistema Integrado para Classificação Automatizada de Serviços em NFS-e com Machine Learning" na Sessão Demos e Aplicações do SBBD 2025

Your SBBD 2025 - DEMOS paper 2[redacted]6

1 mensagem

JEMS <jems@sbc.org.br>
Responder a: lcjunior@utfpr.edu.br
Para: tarcisioparaiso@gmail.com
Cc: Thiago Mendes <tsoutom@gmail.com>

Dear Tarcísio Paraiso Farias:

Congratulations - your paper "Nota Conforme: Sistema Integrado para Classificação Automatizada de Serviços em NFS-e com Machine Learning" for SBBD 2025 - DEMOS has been accepted.

To complete the publication process, please ensure that you submit the following items by July 25, 2025, through the JEMS system:

a) Final Paper Version: Please include the link to the GitHub repository of the tool associated with your paper.

b) Publication Authorization Form:
https://drive.google.com/file/d/1QejBqnApI6R4pcoOzk_2cXk8rHUBbf8S/view

c) Receipt of Author Registration

The reviews are below or can be found at
https://jems.sbc.org.br/PaperShow.cgi?m=2[redacted]6

Best Regards,
Luiz Celso Gomes Jr
SBBD Demos Chair