AVALIAÇÃO DE MODELO DE MACHINE LEARNING PARA CLASSIFICAÇÃO DE INFORMAÇÕES SIGILOSAS: UM ESTUDO DOS DADOS DO GOVERNO FEDERAL

EVALUATION OF A MACHINE LEARNING MODEL FOR SENSITIVE DATA CLASSIFICATION: A STUDY ON FEDERAL GOVERNMENT DATA

Rafaella Weiss Siqueira Costa
rwsc@discente.ifpe.edu.br
Guilherme José de Carvalho Cavalcanti
guilherme.cavalcanti@belojardim.ifpe.edu.br

RESUMO

Nos últimos anos, a Administração Pública tem avançado fortemente na modernização dos sistemas governamentais alinhada à política de Governo Digital. Esse movimento incentiva o uso de novas tecnologias disruptivas, como automações e aprendizado de máquina, que possam contribuir para eficiência dos processos administrativos e para qualidade do serviço público. Nesse contexto, um dos temas centrais é o tratamento de dados, que deve estar em conformidade com a Lei de Acesso à Informação (LAI) e Lei Geral de Proteção de Dados (LGPD). Além disso, é necessária a rigorosa observação de tratativas de segurança da informação na proteção de dados sensíveis. O presente artigo teve como objetivo a avaliação de um modelo de machine learning para classificação de informações sigilosas com base em dados históricos do Governo Federal. O modelo Random Forest apresentou bom desempenho na classificação de informações sigilosas, atingindo 94,39% de acurácia. Outras métricas, como a Precisão-Recall (PRC), foram analisadas. No entanto, desafios no desbalanceamento de classes e na diferenciação entre categorias de sigilo foram identificados, destacando a necessidade de refinamento futuro na modelagem, de coleta de um maior volume representativo de amostras e seleção de características.

Palavras-chave: Governo Digital; Classificação de informações; Machine Learning.

ABSTRACT

In recent years, Public Administration has made significant progress in modernizing government systems in alignment with the Digital Government policy. This movement encourages the adoption of disruptive technologies, such as automation and machine learning, which can enhance administrative processes' efficiency and improve public service quality. Data processing has become a central issue in this context, requiring compliance with the Access to Information Law (LAI) and the General Data Protection Law (LGPD).

Instituto Federal de Pernambuco. Campus Belo Jardim. Curso de Engenharia de Software. 12 de fevereiro de 2025.

Furthermore, strict adherence to information security protocols is essential to safeguard sensitive data. This article aims to evaluate a machine learning model for classifying confidential information based on historical data from the Federal Government. To this end, the study is divided into three sections: the first presents the justification and motivation for the work, along with the theoretical framework, considering the advancements in the use of ICTs in public services; the second section addresses the experimental structure, detailing the methodology used for data collection and processing, as well as the experiments conducted; finally, the results, challenges faced, and future work are discussed. The Random Forest model demonstrated good performance in classifying confidential information, achieving an accuracy of 94.39%. Other metrics, such as the Precision-Recall Curve (PRC), were also analyzed. However, challenges related to class imbalance and the differentiation between confidentiality categories were identified, highlighting the need for future refinements in modeling, the collection of a more representative volume of samples, and feature selection.

Keywords: Digital Government; Data Classification; Machine Learning.

1 INTRODUÇÃO

Nos últimos anos, a Administração Pública tem avançado fortemente na modernização dos sistemas governamentais alinhada à política de Governo Digital. Esse movimento incentiva o uso de novas tecnologias disruptivas, como automações e aprendizado de máquina, que possam contribuir para eficiência dos processos administrativos e para qualidade do serviço público. Nesse contexto, um dos temas centrais é o tratamento de dados, que deve estar em conformidade com a Lei de Acesso à Informação (LAI) e Lei Geral de Proteção de Dados (LGPD). Além disso, é necessária a rigorosa observação de tratativas de segurança da informação na proteção de dados sensíveis.

A Estratégia Nacional do Governo Digital (ENGD) está prevista na Lei nº 14.129, de 29 de março de 2021 (Lei do Governo Digital). Essa política se dedica ao regimento de princípios, regras e instrumentos para o aumento da eficiência da administração pública, especialmente por meio da desburocratização, da inovação, da transformação digital e da participação do cidadão (BRASIL, 2024). O uso de novas tecnologias está expressamente previsto em diversas legislações brasileiras recentes, como os normativos do Conselho Nacional de Justiça sobre a Plataforma Digital do Poder Judiciário - PDPJ e a Portaria SGD/MGI Nº 4.248/2024. A aderência de tecnologia da informação e comunicação como agente transformador do serviço público é crescente no âmbito da administração pública.

Conforme o estudo de Cristóvam, Saikali e Sousa (2020), a prestação de serviços públicos pelo meio digital permite a interação entre fatores humanos e organizacionais à tecnologia de informação (aparatos tecnológicos), que são responsáveis pela captura,

armazenamento, transmissão e manipulação de informação, fomentando a prática de uma gestão pública típica do Governo Digital. Inclusive, há abertura explícita ao uso de Inteligência Artificial em registro de dados e pedidos administrativos, facilitando a tutela administrativa, contribuindo para a maior eficácia dos precedentes administrativos.

O emprego crescente de tecnologias da informação na administração pública tem como objetivos principais a eficiência e potencialização das rotinas administrativas e acesso à informação pelos cidadãos (CRISTÓVAM; SAIKALI; SOUSA, 2020). Há de se garantir, ainda, a observância da segurança das informações classificadas. A implementação das TICs no âmbito governamental não se dá de forma simples, visto que deve seguir legislações e processos burocráticos específicos. No entanto, a aplicação das TICs tem apresentado sucesso na garantia de direitos sociais, como a democratização da informação.

Neste contexto, o objetivo principal deste trabalho é realizar uma avaliação de um modelo de machine learning no apoio à classificação de informações sigilosas a partir de dados do Governo Federal. Através da aplicação de modelo aprendizado de máquina, busca-se auxiliar na classificação de documentos e dados sensíveis conforme normas e dispositivos legais, reduzindo o risco de acesso não autorizado e potencializando a proteção da informação. Além disso, a automação do processo de classificação facilita a conformidade com requisitos de governança, garantindo que a manipulação de dados sigilosos esteja alinhada com políticas de segurança e melhores práticas, o que fortalece a transparência e a eficiência nas atividades administrativas. Ademais, o uso de classificação automática se alinha à promoção de uma gestão mais ágil e da eficiência no serviço público.

Este estudo está dividido em três seções: primeiro, apresenta-se a justificativa e motivação do trabalho e fundamentação teórica sobre as movimentações no uso das TICs no serviço público, leis ou regulamentações que apoiam a temática e conceitos do domínio de aprendizado de máquina; a segunda seção aborda a estrutura da experimentação, detalhando a metodologia utilizada para coleta e tratamento de informações, assim como experimentos realizados; por fim, discute-se sobre os resultados obtidos, desafios enfrentados e trabalhos futuros.

Foram utilizados conjuntos de dados de informações classificadas advindos dos Ministérios da Justiça, da Saúde e da Fazenda, do período de 2019 a 2023. Os dados passaram por tratamento e pré-processamento; estas etapas, bem como o treinamento do modelo, são descritas na seção de metodologia. Entre os resultados e discussão está a

avaliação do modelo, considerando métricas indicadas pela literatura que levam em consideração fatores como dimensionalidade e balanceamento das classes, e suas limitações. O modelo apresentou performance similar para as três classes de sigilo nos relatórios de classificação e curva Precisão-Recall, mas desperta questionamento sobre as características do conjunto de dados, como desbalanceamento de classes e quantidade de amostras representativas, que podem influenciar no viés e classificação.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Governo Digital

A introdução de serviços eletrônicos pelas administrações públicas datam do fim da década de 1980, mas a criação de políticas públicas que estimulam o uso de TICs se dá a partir da década de 90, período de forte efeito da globalização. O governo eletrônico se posiciona a partir de duas perspectivas: de forma estrita, governo eletrônico reside no uso das Tecnologias da Informação – TICS pelos governos; sob uma visão ampla, o governo eletrônico é o uso de TICs para garantir acesso a informações e serviços públicos e transformar governos (VIANA, 2021).

A utilização de tecnologias da informação e comunicação está cada vez mais integrada à rotina das entidades da administração pública, de forma a melhorar a eficiência e gestão dos processos administrativos. Conforme o estudo de Gouveia (2004), o uso das TICs pode considerar a separação por três grandes áreas de intervenção:

- E-administração: melhoria dos processos associados ao funcionamento do poder político e da Administração Pública;
- E-cidadãos e e-serviços: interligação entre cidadãos e empresas, por oferta de valor e serviços;
- E-sociedade: desenvolvimento e construção de interações externas ao poder político e Administração Pública. Normalmente associados a questões de participação pública e cidadania.

A transformação digital no serviço público é apoiada em diversas legislações que se orientam pela política de Governo Digital. A Estratégia Nacional de Governo Digital (ENGD) está prevista na Lei nº 14.129, de 29 de março de 2021 (Lei do Governo Digital) e foi

elaborada, sob a coordenação da Secretaria de Governo Digital, a partir de um amplo processo participativo.

Em Brasil (2024), a ENGD é descrita como conjunto de recomendações que objetivam articular e direcionar iniciativas de governo digital em todos os entes federados e tem como propósitos:

- A promoção da transformação digital no serviço público que vise aprimorar a eficiência, a transparência, a acessibilidade e o impacto positivo dos serviços governamentais;
- O fortalecimento da participação dos cidadãos e impulso à inovação tecnológica;
- A criação de uma administração pública mais moderna e ágil.

A ENGD lista dez objetivos específicos que tratam sobre temas como segurança, infraestrutura, transparência e participação, inteligência de dados, entre outros. Dentre as recomendações postuladas, há o estímulo da adoção de tecnologias disruptivas como ciência de dados para tomada de decisão das políticas públicas e e personalização dos serviços. Além do mais, a inteligência de dados deve servir para catalogar e promover a descoberta e reuso de dados.

2.2 Lei de Acesso à Informação e informações classificadas

A Lei de Acesso à Informação (LAI) - Lei N° 12.527, de 18 de novembro de 2011 - é uma lei que regula e garante o direito fundamental de acesso à informação pelo público, observando diretrizes como a observância da publicidade como preceito geral e do sigilo como exceção; a divulgação de informações de interesse público independentemente de solicitações; a utilização de meios de comunicação viabilizados pela tecnologia da informação; o fomento ao desenvolvimento da cultura da transparência na administração pública; e o desenvolvimento do controle social da administração pública. A esta legislação estão subordinados, conforme texto da lei:

> I - os órgãos públicos integrantes da administração direta dos Poderes Executivo, Legislativo, incluindo as Cortes de Contas, e Judiciário e do Ministério Público;

> II - as autarquias, as fundações públicas, as empresas públicas, as sociedades de economia mista e demais entidades controladas direta ou indiretamente pela União, Estados, Distrito Federal e Municípios.

A LAI caracteriza informação como todos os dados, processados ou não, que podem ser utilizados para produção e transmissão de conhecimento, contidos em qualquer meio, suporte ou formato. Além disso, a lei também descreve sobre a qualidade e o sigilo da informação. Informações imprescindíveis para a segurança da sociedade e do Estado são classificadas como sigilosas, e são submetidas temporariamente à restrição de acesso público. O artigo 23 define que serão classificadas informações que possam:

> I - pôr em risco a defesa e a soberania nacionais ou a integridade do território nacional;
>
> II - prejudicar ou pôr em risco a condução de negociações ou as relações internacionais do País, ou as que tenham sido fornecidas em caráter sigiloso por outros Estados e organismos internacionais;
>
> III - pôr em risco a vida, a segurança ou a saúde da população;
>
> IV - oferecer elevado risco à estabilidade financeira, econômica ou monetária do País;
>
> V - prejudicar ou causar risco a planos ou operações estratégicos das Forças Armadas;
>
> VI - prejudicar ou causar risco a projetos de pesquisa e desenvolvimento científico ou tecnológico, assim como a sistemas, bens, instalações ou áreas de interesse estratégico nacional;
>
> VII - pôr em risco a segurança de instituições ou de altas autoridades nacionais ou estrangeiras e seus familiares; ou
>
> VIII - comprometer atividades de inteligência, bem como de investigação ou fiscalização em andamento, relacionadas com a prevenção ou repressão de infrações.

O Artigo 24 define que informações com teor sensível poderão ser classificadas em três níveis: ultrassecreta, secreta e reservada, cada uma com um prazo máximo de restrição de acesso — respectivamente, 25, 15 e 5 anos, contando a partir da data de produção da informação. Em que pese sejam dados públicos, a própria LAI restringe o acesso a esses por um tempo determinado, a depender do grau de sigilo, demonstrando que tal enquadramento não impede que haja limitação do acesso pelo público geral (CRISTÓVAM; HAHN, 2020).

De acordo com o Artigo 30 da LAI, anualmente, a autoridade máxima de cada órgão ou entidade deverá publicar em sítio eletrônico o rol das informações que tenham sido

desclassificadas e classificadas, seus graus de sigilo, bem como relatório que demonstra o quantitativo de pedidos de informação no período.

2.3 Aprendizado de Máquina para tarefas de classificação

Aprendizado de Máquina é uma subárea da inteligência artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática (MONARD; BARANAUSKAS, 2003).

As técnicas de aprendizado de máquina estão sendo utilizadas em muitos domínios do conhecimento para construção de sistemas que resolvam dores específicas e ofereçam um melhor serviço aos seus usuários. Esses sistemas tornam-se cada vez mais robustos, principalmente pelo avanço da tecnologia e o grande volume de dados disponível que serve de base para o aprendizado.

Considerando a hierarquia do aprendizado, ele é dividido em três tipos principais: supervisionado, não-supervisionado e por reforço.

No aprendizado supervisionado um conjunto de exemplos de treinamento com rótulos previamente definidos apoiam a decisão do algoritmo para novos dados ainda não rotulados. Conforme Monard e Baranauskas (2003), para rótulos de classe discretos, esse problema é conhecido como classificação e para valores contínuos como regressão. De acordo com Rosa et al. (2023), a tarefa de classificação, a qual é objeto de estudo desse trabalho, realizada por aprendizado de máquina supervisionado, consiste na análise do conjunto de registros fornecidos – onde cada registro apresenta a indicação da classe pertencente – e aprende com esses registros para que, na fase posterior, classifique automaticamente um novo registro.

No aprendizado não-supervisionado, o algoritmo deverá prever a classe alvo dos exemplos sem o apoio de rótulos. Ao identificar similaridades dos seus atributos, são formados agrupamentos ou clusters. Monard e Baranauskas (2003) e Ludermir (2021) concordam que, em geral, após a determinação dos agrupamentos, é necessária uma análise que determine o significado desses grupos no contexto problema sendo analisado.

No Aprendizado por Reforço, o algoritmo não recebe a resposta correta mas recebe um sinal de reforço, de recompensa ou punição. O algoritmo faz uma hipótese baseado nos exemplos e determina se essa hipótese foi boa ou ruim (LUDERMIR, 2021).

2.3 Terminologias comuns em aprendizado de máquina

2.1.1 Hiperparâmetros

Hiperparâmetros são o conjunto de configurações de um modelo que pode ser utilizado para ajustar e melhorar sua performance (BERGSTRA; BENGIO, 2012). Segundo Bergstra e Bengio (2012), a configuração correta estaria mais relacionada com os resultados durante o experimento do que com a teoria. Ou seja, em várias das ocasiões, a seleção de hiperparâmetros ótimos se baseia em combinações e testes (Busca Aleatória) para avaliar o desempenho do modelo até um resultado que seja satisfatório para o pesquisador.

No entanto, para outros autores, como Patterson e Gibson (2017), ajustar o modelo seguindo os resultados experimentais leva a um problema comum de overfitting (sobreajuste) e, para minimizar este incoveniente, é necessário o emprego de alguma técnica que varie os conjuntos de dados de entrada, como é a validação cruzada utilizada neste trabalho.

2.1.1 Overfitting e Underfitting

Overfitting (sobreajuste) é um problema encontrado em aprendizado de máquina quando o modelo se ajusta excessivamente aos dados de treinamento, não generalizando bem para dados de teste ou padrões diferentes dos vistos em treino (HAYKIN, 2017; REIS, 2018). O overfitting pode acontecer em um modelo com configurações muito complexas, dimensionalidade ou tamanho não satisfatório do conjunto de dados, entre outros fatores.

Underfitting (subajuste) acontece quando o modelo não consegue aprender nenhum padrão nos dados ou fazer correlações. O underfitting pode aparecer em modelos muito simples, performando mal tanto em treino quanto em teste. Esse problema é derivado de fatores como baixa dimensionalidade e seleção de características (ALIFERIS; SIMON, 2024).

3 METODOLOGIA

O objeto de estudo deste trabalho é a avaliação de um modelo de machine learning para para a classificação de informações sigilosas com base em dados históricos do Governo Federal. O foco está em apoiar a conformidade com legislações como a Lei de Acesso à Informação (LAI) e a Lei Geral de Proteção de Dados (LGPD), além de melhorar a segurança da informação, evitando falhas humanas que resultam em vazamento de informações sigilosas e reduzindo a probabilidade de subjetividade no processo de classificação.

A metodologia adotada neste estudo seguiu quatro etapas principais: (1) Coleta e preparação dos dados, (2) Seleção de características, (3) Treinamento e validação do modelo, e (4) Avaliação dos resultados. A seguir, cada uma dessas etapas será detalhada

Para este trabalho, foram selecionadas bases de dados de informações classificadas pelo Ministério da Justiça, Ministério da Saúde e Ministério da Fazenda nos anos de 2019 a 2023. Embora disponibilizados no período citado, alguns dos dados levantados foram produzidos em datas anteriores. O rol de informações classificadas e desclassificadas, como já posto anteriormente neste trabalho, deve ser anualmente divulgado nos sites dos órgãos e entidades públicas, de acordo com a Lei de Acesso à Informação.

O software utilizado para o tratamento foi o Google Colab com a linguagem de programação Python e as bibliotecas Scikit-Learn, Numpy, Matplotlib, entre outras. A linguagem Python é amplamente utilizada para trabalhos de aprendizagem de máquina pela sua versatilidade, ampla variedade de bibliotecas e frameworks para construção de modelos e documentação detalhada.

3.1 Seleção de características

Inicialmente, as três bases de dados dos Ministérios da Justiça, da Saúde e da Fazenda, conforme pode ser visto na Figura 1, Figura 2 e Figura 3, respectivamente, possuíam diversas colunas que não eram de interesse para o objeto de estudo. Na etapa de preparação dos dados foi realizada a exclusão de colunas, bem como o tratamento de valores faltantes e a exclusão de valores nulos.

Cada base foi, previamente, tratada de forma separada. Posteriormente, as bases foram combinadas, formando um único dataset que soma aproximadamente 4.600 linhas e 5 colunas, demonstrado na Figura 4.

As colunas selecionadas dispõem informações sobre o dispositivo legal associado, o grau de sigilo da informação, tipo do assunto, o assunto tratado e categoria, classificadas de acordo com a fundamentação dos decretos 7.845 e 7.724. A classe alvo para classificação neste trabalho é o atributo que determina o grau de sigilo da informação, seja secreto, ultrassecreto ou reservado.

Algumas colunas foram descartadas pois não adicionavam informações relevantes para construção do modelo, como N. de ordem; códigos identificadores e de indexação; CIDIC; datas de produção e classificação; unidade custodiante e orgão.

Outras colunas tratavam da mesma informação, mas exibiam títulos diferentes nas bases de dados como, por exemplo, “Prazo de restrição de acesso" (Figura 3) com “Prazo de Classificação" (Figura 1) e “Categoria (VCGE)” (Figura 3) e “Categoria” (Figura 2). Estas colunas foram foram renomeadas para combinação das bases de dados e geração do dataset final.

Figura 1 – Dados do Ministério da Justiça

	Nº Ord	Código de Indexação de Documento que contém Informação Classificada – CIDIC	Categoria[1]	Dispositivo Legal[2]	Data de Produção	Data de Classificação	Prazo de Classificação	Validade da Classificação	Unidade Custodiante	Grau de Sigilo	Tipo de Assunto
0	1	08001.001468/2018-51.S.05.04/04/2018.03/04/2033.N	Defesa e Segurança	III	04/04/2018	19/04/2018	15 anos	03/04/2033	CPADS	secreto	Saúde Pública
1	2	08200.003733/2020-11.S.05.02/03/2020.04/03/2035.N	Defesa e Segurança	VIII	02/03/2020	04/03/2020	15 anos	04/03/2035	PF	secreto	Inteligência e Investigação
2	3	08123.000415/2013-21.S.05.22/07/2008.21/07/2023.N	Defesa e Segurança	VIII	22/07/2008	29/01/2013	15 anos	22/07/2023	PF	secreto	Inteligência e Investigação
3	4	08064.001653/2013-04.S.05.14/06/2012.13/06/2027.S	Defesa e Segurança	VIII	14/06/2012	14/06/2012	15 anos	14/06/2027	PF	secreto	Inteligência e Investigação
4	5	08064.001651/2013-15.S.05.02/07/2012.01/07/2027.S	Defesa e Segurança	VIII	02/07/2012	02/07/2012	15 anos	02/07/2027	PF	secreto	Inteligência e Investigação
...	...	...	...	...	...	...	...	...	...	...	...
4249	2090	08016.015058/2012-32.U.05.12/09/2012.11/09/2037.N	Defesa e Segurança	III	12/09/2012	12/04/2013	25 anos	12/09/2037	DEPEN	ultrassecreto	Saúde Pública
4250	2091	08016.015749/2012-36.U.05.27/09/2012.26/09/2037.N	Defesa e Segurança	III	27/09/2012	12/04/2013	25 anos	27/09/2037	DEPEN	ultrassecreto	Saúde Pública
4251	2092	08016.016544/2012-78.U.05.09/10/2012.08/10/2037.N	Defesa e Segurança	III	09/10/2012	12/04/2013	25 anos	09/10/2037	DEPEN	ultrassecreto	Saúde Pública

Fonte: Ministério da Justiça (2024).

Figura 2 – Dados do Ministério da Saúde

	Unnamed: 0	CIDIC	CATEGORIA (15)	GRAU SIGILO	DISPOSITIVO LEGAL	DATA DA PRODUÇÃO	DATA DA CLASSIFICAÇÃO	PRAZO DA CLASSIFICAÇÃO	ASSUNTO
0	1	250007.25000.131860/2012-67.S.15.12/11/2012.12...	SAUDE	SECRETO	Inciso VI, art. 23, Lei n. 12527/2011	2012-11-12	2012-11-12	2027-11-12	Processo Administrativo contendo Termo de Comp...
1	2	250007.25000.049824/2011-70.S.15.19/05/2011.19...	SAUDE	SECRETO	Inciso VI, art. 23, Lei n. 12527/2011	2011-05-19	2011-05-19	2026-05-19	Processo Administrativo contendo Termo de Comp...
2	3	250007.25000.191535/2012-53.S.15.12/11/2012.12...	SAUDE	SECRETO	Inciso VI, art. 23, Lei n. 12527/2011	2012-11-12	2012-11-12	2027-11-12	Processo Administrativo contendo Termo de Comp...
3	4	250007.25000.191511/2012-02.S.15.12/11/2012.12...	SAUDE	SECRETO	Inciso VI, art. 23, Lei n. 12527/2011	2012-11-12	2012-11-12	2027-11-12	Processo Administrativo contendo Termo de Comp...
4	5	250007.25000.136099/2012-50.S.15.12/11/2012.12...	SAUDE	SECRETO	Inciso VI, art. 23, Lei n. 12527/2011	2012-11-12	2012-11-12	2027-11-12	Processo Administrativo contendo Termo de Comp...
...	...	...	...	...	...	...	...	...	...
153	154	250007.25000095575/2018-61.R.15.02/01/2018.01/...	SAUDE	RESERVADO	Incisos III, IV e VII do artigo 23 da lei 12.527.	2018-01-02	2018-06-01	2023-06-01	Controle de Estoque de Insumos Estratégicos pa...
154	155	250007.25000095620/2018-87.R.15.02/01/2018.01/...	SAUDE	RESERVADO	Incisos III, IV e VII do artigo 23 da lei 12.527.	2018-01-02	2018-06-01	2023-06-01	Plano de Demandas de Insumos Estratégicos para...
155	156	25000.142407/2020-96.R.15.08/10/2020.08/10/2025.N	SAUDE	RESERVADO	Inciso III, VI e VIII - Art. 23 da lei 12527	2020-10-08	2020-12-11	2025-10-08	Plano Nacional de Imunização para CORONAVIRUS
156	157	25000169275/2022-40.R.15.02/10/2022.02/10/2025.N	SAUDE	RESERVADO	Inciso III, §1° - Art. 24 da lei 12527	2020-10-02	2020-10-02	2025-10-02	Análise preparatória acerca de aquisição de in...
157	158	25000.009084/2021-19 R.15.08/2021.08.02.2026.N	SAUDE	RESERVADO	Inciso III - Art. 24 da lei 12527	2021-02-08	2021-02-08	2025-02-08	Proposta de cooperação Internacional para prod...

Fonte: Ministério da Saúde (2024).

Figura 3 – Dados do Ministério da Fazenda

	NUP	Grau de Sigilo	Categoria (VCGE)	Data de Produção	Data de Classificação	Prazo de Restrição de Acesso	Data da Desclassificação	Indicação de Reclassificação	Fundamentação Legal para Classificação	Assunto do Documento\n(breve resumo)	Órgão
0	10167.720001/2022-49	Reservado	06 - Economia e Finanças	2021-04-27 00:00:00	2022-04-13 00:00:00	5 anos	2026-04-27 00:00:00	Não	Art. 23, inciso VIII da Lei 12527/2011	Auditoria	RFB
1	10265.044038/2020-99	Reservado	06 - Economia e Finanças	2020-10-01 00:00:00	2022-10-10 00:00:00	5 anos	2025-01-01 00:00:00	Não	Art. 23, inciso VIII da Lei 12527/2011	Planejamento	RFB
2	10265.094663/2020-81	Reservado	06 - Economia e Finanças	2020-03-01 00:00:00	2022-10-10 00:00:00	5 anos	2025-03-01 00:00:00	Não	Art. 23, inciso VIII da Lei 12527/2011	Planejamento	RFB
3	10265.248279/2022-77	Reservado	06 - Economia e Finanças	2021-10-01 00:00:00	2022-10-10 00:00:00	5 anos	2026-10-01 00:00:00	Não	Art. 23, inciso VIII da Lei 12527/2011	Planejamento	RFB
4	13031.172164/2023-17	Reservado	06 - Economia e Finanças	2023-06-26 00:00:00	2023-06-26 00:00:00	5 anos	2028-06-26 00:00:00	Não	Art. 23, inciso VIII da Lei 12527/2011	Gestão do Crédito Tributário e do Direito Cred...	RFB
...	...	...	...	...	...	...	...	...	...	...	...
182	12177.100256/2020-35	Reservado	06 - Economia e Finanças	2020-07-01 00:00:00	2020-08-28 00:00:00	5 anos	2025-06-30 00:00:00	Não	Artigo 25, inciso V, do Decreto nº 7.724, de 1...	Cálculos (10163787 e 10169912) e modelos (1016...	SPE
183	12177.100267.2019-81	Reservado	06 - Economia e Finanças	2019-06-27 00:00:00	2019-06-27 00:00:00	5 anos	2024-06-27 00:00:00	Não	Lei 12.527/2011, Art. 23, IV.	Nota Técnica	SPE
184	12177.100273/2020-	Reservado	06 - Economia e	2020-07-31	2020-09-11	5 anos	2025-09-10 00:00:00	Não	Artigo 25, inciso V, do Decreto nº	Nota Técnica "Impactos	SPE

Fonte: Ministério da Fazenda (2024).

Figura 4 – Dataset final após primeiro pré-processamento e combinação das bases

	Categoria	Dispositivo Legal	Grau de Sigilo	Tipo de Assunto	Assunto
0	Defesa e Segurança	III	secreto	Saúde Pública	Processo Administrativo contendo Termo de Comp...
1	Defesa e Segurança	VIII	secreto	Inteligência e Investigação	Processo Administrativo contendo Termo de Comp...
2	Defesa e Segurança	VIII	secreto	Inteligência e Investigação	Processo Administrativo contendo Termo de Comp...
3	Defesa e Segurança	VIII	secreto	Inteligência e Investigação	Processo Administrativo contendo Termo de Comp...
4	Defesa e Segurança	VIII	secreto	Inteligência e Investigação	Processo Administrativo contendo Termo de Comp...
...	...	...	...	...	...
4602	Economia e Finanças	V	reservado	Forças Armadas	Cálculos (10163787 e 10169912) e modelos (1016...
4603	Economia e Finanças	IV	reservado	Estabilidade Financeira	Nota Técnica
4604	Economia e Finanças	V	reservado	Forças Armadas	Nota Técnica "Impactos Iniciais da Introdução ...
4605	Economia e Finanças	V	reservado	Forças Armadas	Nota Técnica "Impactos redistributivos da CBS:...
4606	Defesa e Segurança	VIII	reservado	Inteligência e Investigação	Processo Administrativo contendo Termo de Comp...

4607 rows × 5 columns

Fonte: Elaboração do autor.

3.1.1 Características dos atributos no conjunto de dados

Aqui pretende-se explicar melhor as características de cada um dos atributos (colunas) no conjunto de dados após realizada a combinação das bases, abordando o tipo do dado e sua relevância para construção do modelo.

Tabela 1 – Informações dos atributos do conjunto de dados

Atributo	Tipo	Descrição
Categoria	Object	Categoria temática do documento.
Dispositivo Legal	Object	Dispositivo legal que fundamenta a classificação do documento (Inciso do artigo 23 da LAI).
Grau de Sigilo	Object	Nível de sigilo do documento (reservado, secreto, ultrassecreto).
Tipo de Assunto	Object	Tipo de assunto relacionado ao dispositivo legal, de acordo com a LAI.
Assunto	Object	Breve resumo do assunto do documento.

Fonte: Elaboração do autor.

Para o conjunto de dados combinados dos foram selecionadas cinco colunas para utilização como características na construção do modelo, estas estão apresentadas na Tabela 1. Todas as colunas selecionadas originalmente eram do tipo String. Com a biblioteca Pandas, utilizada para criação dos dataframes, dados do tipo String são tratados como tipo Object.

A coluna “Categoria” descreve a temática do documento, como Defesa e Segurança, Economia e Finanças, Saúde. Certos temas podem estar associados a um grau de sigilo maior.

O “Dispositivo Legal” é o inciso que fundamenta a classificação da informação, de acordo com o contexto associado ao artigo 23 da LAI, esta coluna pode auxiliar o modelo a correlacionar padrões legais com diferentes níveis de confidencialidade.

A coluna “Grau de Sigilo” serve como referência para classificação, sendo a variável alvo, a classe que o modelo deve prever com base nos demais atributos.

As colunas “Tipo de Assunto” e “Assunto” trazem informações sobre as matérias tratadas no documento classificado, fornecendo contexto adicional que pode indicar a necessidade de sigilo com base no tema tratado.

3.2 Balanceamento de classes

Após o tratamento inicial dos dados e combinação das bases, notou-se um desbalanceamento visível entre as classes alvo. O desbalanceamento de classes é um dos fatores que podem impactar na performance do modelo, por influenciar diretamente o viés e, consequentemente, a classificação correta.

Conforme Castro e Braga (2011) e Chawla et al. (2002), o problema das classes desbalanceadas é comum e aparece com frequência em muitos dos domínios de estudo e aplicação de aprendizado de máquina, tendo sido interesse crescente por outros pesquisadores nos últimos anos. Existem esforços para lidar com este problema em áreas como detecção de chamadas telefônicas fraudulentas, gestão de telecomunicações, classificação de texto, detecção de vazamentos de óleo em imagens de satélite, entre outras.

Um dos aspectos fundamentais em problemas de classificação é a disparidade na distribuição de classes, que surge principalmente em situações onde informações associadas a determinadas classes são mais difíceis de serem obtidas. Pode-se observar esse comportamento, por exemplo, em um estudo sobre uma doença rara em uma dada população (CASTRO; BRAGA, 2011).

Não é incomum, segundo He e Garcia (2009) o desbalanceamento entre classes, sendo inclusive muito presente na ordem de 100:1, 1.000:1 e até 10.000:1, onde uma classe supera as outras. Os autores trazem o exemplo do campo da biomedicina, em eventos em pacientes com tumores malignos, onde o número de amostras positivas (pacientes com câncer) é

facilmente superado pelo número de amostras negativas (pacientes saudáveis). Nota-se que esse desbalanceamento pode envolver conjuntos de dados onde existem multiclasses em diversos problemas do mundo real, como é apresentado neste trabalho.

Neste trabalho, observa-se uma grande diferença em número entre informações classificadas como reservadas (maioria) daquelas classificadas como secretas ou ultrassecretas (minoria). Esta discrepância pode estar associada a fatores como critérios de classificação, tempo menor de restrição para certas informações, menor risco para matérias de segurança nacional, entre outros. A predominância de informações reservadas também reflete um equilíbrio entre proteção e transparência, garantindo que apenas dados extremamente sensíveis permaneçam restritos por um período de tempo maior.

Conforme Farquad e Bose (2012), os métodos para lidar com dados desbalanceados incluem o redimensionamento do conjunto de treinamento, aplicando a superamostragem de amostras de classes minoritárias (oversampling) ou a redução do tamanho das amostras de classes majoritárias (undersampling), o ajuste dos custos de classificação incorreta e o aprendizado baseado em reconhecimento.

Com o objetivo de balancear as classes do dataset, avaliou-se três técnicas de balanceamento:

- Oversampling: esta técnica visa aumentar a quantidade de dados da classe minoritária.
- Undersampling: o objetivo da técnica é diminuir a quantidade de dados da classe majoritária.
- SMOTE (*Synthetic Minority Oversampling Technique*): aumenta a quantidade de dados das classes minoritárias através da geração de exemplos sintéticos.

No conjunto de dados construído para o estudo, mais de 94% dos dados eram da classe majoritária, informações classificadas como "reservadas", como apresentado na Figura 5. Como elencado por Castro e Braga (2011) ao revisar a literatura, apesar das técnicas de oversampling e undersampling possuírem o mesmo objetivo, elas introduzem outras características ao conjunto de treinamento que podem dificultar o aprendizado.

Além disso, estas técnicas podem excluir aspectos importantes que fazem a diferenciação entre as classes, por exemplo, a técnica de undersampling inclui perda de informação ao

eliminar amostras representativas da classe majoritária. Os autores ainda ressaltam que a escolha de uma abordagem de reamostragem não é uma escolha trivial, inclusive pode ser guiada por outros algoritmos como KNN (K-Nearest Neighbors).

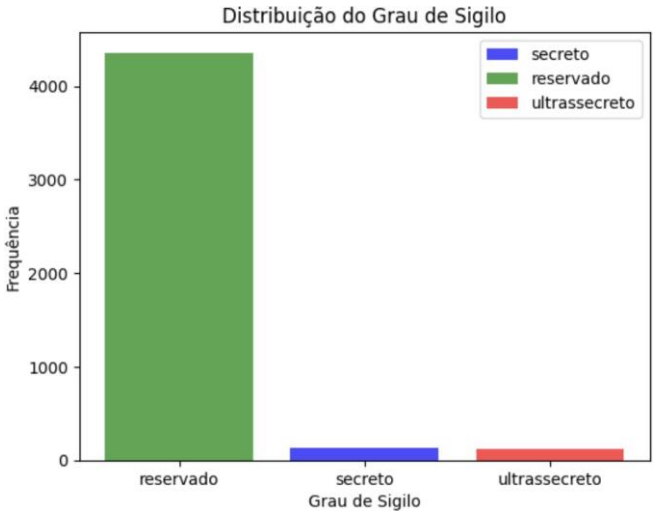
A técnica de reamostragem SMOTE foi selecionada para o balanceamento de classes, equiparando a quantidade de amostras das classes minoritárias (secreto e ultrassecreto) à classe majoritária (reservado) com a criação de amostras sintéticas. Estas amostras sintéticas são baseadas nas semelhanças do espaço de características das amostras das classes minoritárias (HE; GARCIA, 2009).

O SMOTE também possui suas limitações como supergeneralização e variância, demonstradas por He e Garcia (2009). Os estudiosos abordam o fato que o SMOTE cria o mesmo número de amostras sintéticas para cada classe minoritária sem talvez considerar os exemplos próximos ou vizinhos, o que poderia acarretar em sobreposição das classes.

Em um trabalho futuro, o estudo pode ser aprofundado no uso de uma técnica de reamostragem mais avançada. Para este trabalho, o SMOTE é uma alternativa para equilibrar o problema das classes desbalanceadas.

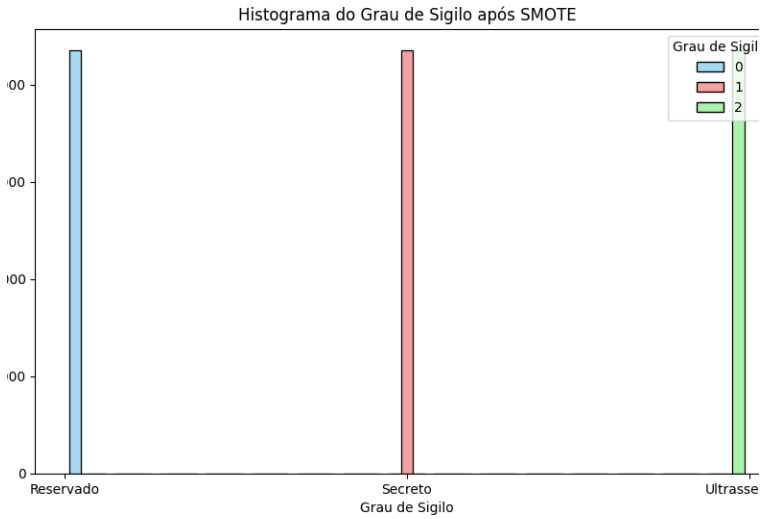
A distribuição das amostras do dataset após a aplicação da técnica SMOTE está demonstrada na Figura 6.

Figura 5 – Distribuição das classes no conjunto de dados inicial



Fonte: Elaboração do autor.

Figura 6 – Distribuição das classes no conjunto de dados após balanceamento com a técnica SMOTE.



Fonte: Elaboração do autor.

3.3 Classificador Random Forest

O algoritmo Random Forest foi selecionado para construção do modelo neste trabalho por ser um algoritmo de aprendizagem supervisionada amplamente estudado e que já obteve desempenho satisfatório em muitos trabalhos, como elencado por Huljanah et al. (2019). O algoritmo Random Forest pode realizar tarefas de classificação e regressão, trabalhando com a construção de várias árvores de decisão baseadas em atributos e características do dataset. É um algoritmo flexível e mais simples.

Conforme Lima et al. (2021), o algoritmo Random Forest ou RF7 se baseia na estratégia de *ensembles*. Utilizando o conceito de redistribuição aleatória dos dados, ele consegue prover diversidade. A biblioteca Scikit-learn proporciona uma excelente ferramenta para isto, que mede a importância das características analisando quantos nós das árvores, que usam um dado atributo, reduzem a impureza geral da floresta. O algoritmo calcula o valor automaticamente para cada atributo após o treinamento e normaliza os resultados para que a soma de todas as importâncias seja igual a 1.

3.4 Validação Cruzada (Cross Validation)

A Validação Cruzada é uma técnica que divide o conjunto de dados em K subamostras ou dobras e treina o algoritmo nestas diferentes partições. A validação cruzada é especialmente utilizada para avaliar a generalização do modelo. Esta técnica pode ser eficiente para melhorar

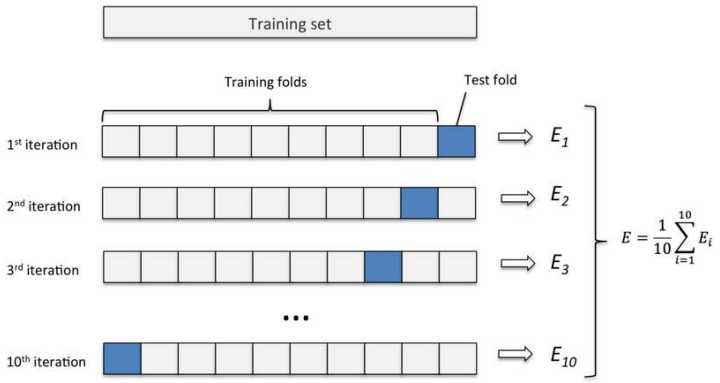
o desempenho em observações não vistas, em ambientes com restrições de dados, pode ser uma ferramenta conveniente para treinar modelos com um conjunto de dados com dimensões menores (MICROSOFT, 2023).

De acordo com Berrar et al. (2019), a validação cruzada pode ser usada para ajustar os hiperparâmetros de modelos estatísticos e de aprendizado de máquina, para evitar o overfitting, comparar algoritmos de aprendizagem e estimar o erro de generalização de um modelo. É uma das técnicas comumente utilizadas para métodos de reamostragem com o objetivo de seleção e avaliação de modelos de aprendizado de máquina.

Uma questão central na aprendizagem supervisionada diz respeito à capacidade de generalização do modelo. Um dos problemas-chave é o overfitting. É muito fácil construir um modelo perfeitamente adaptado ao conjunto de dados de treinamento, mas incapaz de generalizar bem para dados novos e invisíveis (BERRAR et al., 2019).

O modelo deste trabalho foi construído com uma validação cruzada de dez dobras (K=10). Como demonstrado por Kohavi (1995), Borra e Ciaccio (2010) e Kim (2009), a escolha da validação com dez dobras segue um padrão empírico, onde existe um equilíbrio entre o custo computacional e o tamanho de amostras de treino e teste em cada dobra, isso implica uma melhor relação entre viés e variância. Estas dobras são construídas de forma aleatória e, a cada iteração, o conjunto utilizado no treinamento é referente a K-1 subamostras, sendo o K restante utilizado para teste e avaliação da acurácia, apresentado na Figura 7. A acurácia da validação cruzada, por exemplo, é a média de todas as dez medidas registradas

Figura 7 – Divisão dos K-folds na técnica de Validação Cruzada



Fonte: Towards Data Science (2024).

3.5 Acesso ao repositório e modelo

Os códigos de pré-processamento e treinamento do modelo, assim como o conjunto de dados final, foram disponibilizados em repositório público no GitHub (https://github.com/rafxrad/ML_classification_GOV).

O arquivo README fornece orientações de como executá-los no ambiente do Google Colab, incluindo a organização e montagem do Google Drive e as dependências e bibliotecas necessárias para que a execução tenha êxito.

4 RESULTADOS E DISCUSSÃO

Durante os experimentos, o modelo avaliado com o classificador Random Forest atingiu 94.39% de acurácia com 70% das amostras para treinamento e 30% para teste; com 80% dos dados para treino e 20% de reservados como teste, o modelo atingiu 95,11% de acurácia. A métrica de acurácia diz respeito à proporção de previsões corretas realizadas pelo modelo em relação ao total de previsões realizadas, ou seja, a frequência de acertos do modelo.

Para avaliar um modelo de aprendizado de máquina, a avaliação isolada da métrica da acurácia não traduz com exatidão a performance do modelo. Nesse sentido, é necessário avaliar outros quesitos para entender se o modelo satisfaz os resultados iniciais esperados neste trabalho. Considerando Doshi-Vélez e Kim (2017), o problema de uma única métrica, como a avaliação da acurácia na classificação, é que ela é uma descrição incompleta para a maioria das tarefas do mundo real.

De acordo com Erickson e Kitamura (2021), as métricas para classificação podem ser divididas em três grupos principais: binária, multiclasse e multilabel. O presente estudo se baseou em métricas para classificação multiclasse. Classificadores multiclasse selecionam somente uma classe dentre mais de duas classes predefinidas.

4.1 Matriz de confusão

A matriz de confusão é um instrumento de avaliação do modelo de classificação multiclasse, como é o caso da classificação de informações sigilosas. Ela é uma matriz com quatro combinações que permite visualizar a distribuição de predições e valores reais, destacando erros e acertos para cada classe. Nesta pesquisa, foram avaliadas as métricas de acurácia, precisão, recall e F-1 score, descritas na Tabela 2. Estas métricas são fundamentais

para avaliar não somente a taxa de acertos do modelo, como também a capacidade de identificar corretamente cada classe, considerando cenários com desvantagens como o desbalanceamento de classes.

Tabela 2 – Métricas de avaliação

Métrica	Definição	Fórmula
Recall	Proporção de previsões positivas corretas feitas pelo modelo em relação ao total de previsões positivas.	TP/TP + FN
Precisão	Proporção de positivos reais que o modelo conseguiu identificar corretamente.	TP/TP + FP
F1-score	Média harmônica da precisão e do recall	2 x (Precisão x Recall/Precisão + Recall)
Acurácia	A proporção de previsões corretas feitas pelo modelo em relação ao total de previsões realizadas.	Previsões corretas/Total de previsões

Fonte: Elaboração do autor.

A partir do boletim de classificação, gerado pela biblioteca Scikit-learn, é possível observar o desempenho do modelo para tais métricas em todas as três classes (0 - reservado, 1 - secreto, 2 - ultrassecreto).

Ao observar a Figura 8 e Figura 9, podemos analisar o desempenho para as três classes:

Classe 0 (Reservado) - Temos que para classe 0 o modelo tem bom desempenho, atingindo um recall de 98% das classificações. O recall determina que o modelo indicou corretamente a maioria dos dados verdadeiros positivos (TP). A taxa de precisão de 91% indica que poucas amostras foram classificadas de forma errônea com o rótulo "reservado". Este resultado é importante, pois a classe 0 era a classe majoritária no conjunto de dados original e o modelo apresentou uma boa generalização na classificação dessa categoria. Ainda assim, é necessário avaliar o desempenho para a classe majoritária com cautela, pois apresentava inicialmente o maior número de amostras reais em relação às outras classes,

ajustadas com amostras sintéticas, o que pode impactar no viés de classificação ou sobreposição.

Classe 1 (Secreto) - Para esta classe, o modelo apresentou um desempenho mais moderado em relação às outras classes. A taxa de recall foi de 85%, indicando ainda que tenha identificado corretamente a maioria das amostras, uma parcela significativa de falsos positivos foi registrada. No entanto, a precisão ainda teve um valor considerado ótimo, que sugere que o modelo frequentemente classificou corretamente as amostras. O principal desafio aparente para o modelo nessa classe é a diferenciação entre dados rotulados como "secretos" e "ultrassecretos".

Classe 2 (Ultrassecreto) - Para classe 2, o modelo apresentou valores quase perfeitos, sendo a taxa de recall de 100%, tendo classificado corretamente o total de amostras. Dado o alto nível de sigilo associado a esta categoria e também o número de amostras representativas no conjunto de dados original, antes da técnica de reamostragem, este resultado se mostra particularmente relevante para o estudo.

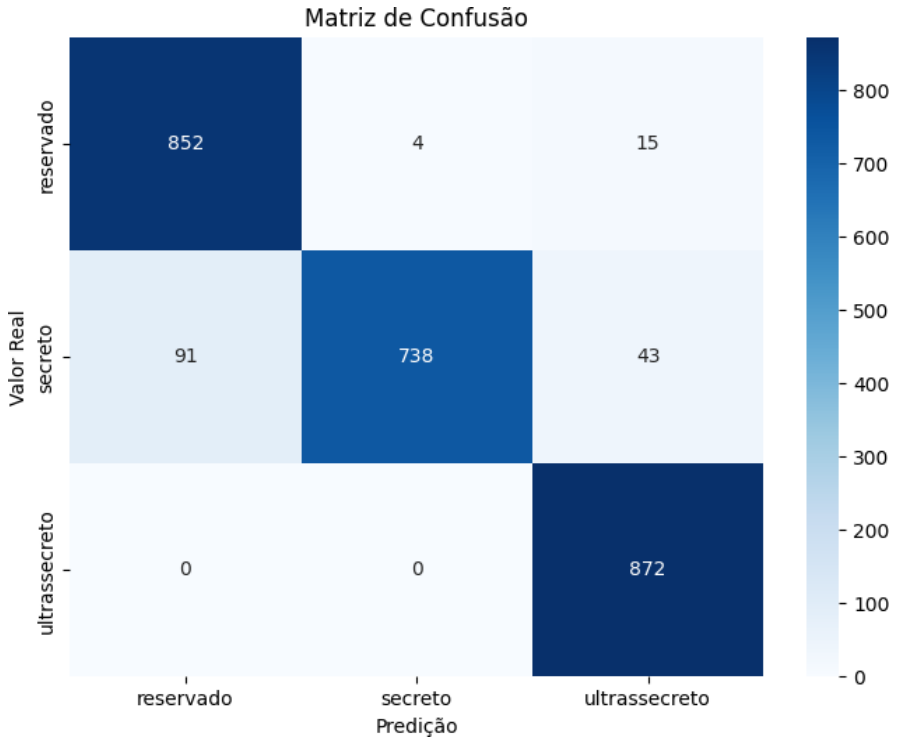
Para as classes minoritárias (1 e 2), vale ressaltar que a maior parte das amostras foi gerada pela técnica SMOTE, de forma sintética. O número de exemplos original da classe 1 e 2, como postulado posteriormente no balanceamento de classes (seção 3.2), indicam a raridade das informações classificadas como muito sensíveis. Nesse sentido, é importante avaliar como estas amostras sintéticas são representativas para classificação nessas classes.

Figura 8 – Boletim de classificação do modelo Random Forest

	precision	recall	f1-score	support
0	0.91	0.98	0.94	1307
1	0.99	0.85	0.92	1307
2	0.94	1.00	0.97	1308
accuracy			0.94	3922
macro avg	0.95	0.94	0.94	3922
weighted avg	0.95	0.94	0.94	3922

Fonte: Elaboração do autor.

Figura 9 – Matriz de confusão do modelo Random Forest



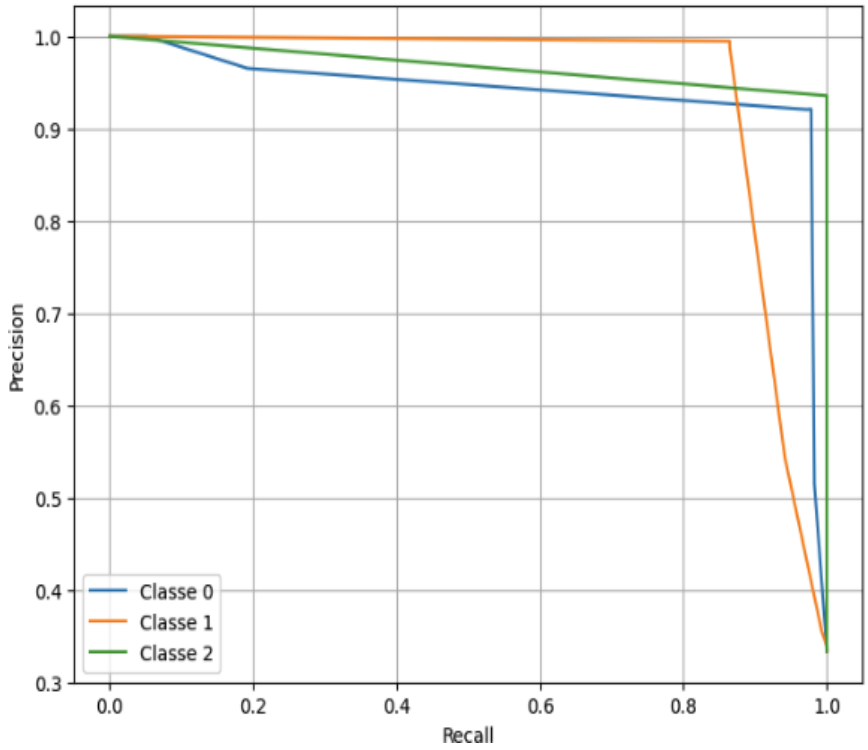
Fonte: Elaboração do autor.

4.2 Curva Precisão-Recall (CPR)

A curva Precisão-Recall (PRC) é um instrumento utilizado para avaliação de classificações baseadas em multiclasse, especialmente em cenários de classificação desbalanceada, pois foca na identificação da classe minoritária. Esta curva mostra a precisão, que é a fração de classificações positivas corretas em relação ao total de valores classificados positivos, frente ao recall, que mede a fração dos exemplos com rótulos positivos frente ao total de verdadeiro positivos.

De acordo com o estudo de Miao e Zhu (2022), para um classificador com bom desempenho, a curva Precisão-Recall deve passar o mais próximo possível do canto superior direito, que indica que o desempenho das observações feitas em teste são melhores. Esta curva geralmente é tortuosa, apresentando variações verticais.

Figura 10 – Curva Precisão-Recall do modelo



Fonte: Elaboração do autor.

Observando a Figura 10, podemos concluir algumas diferenças no desempenho das três classes:

Classe 0 (Reservado) - A classe 0, demonstrada pela linha azul, exibe uma queda gradual de precisão conforme o aumento do recall. O classificador atinge um equilíbrio interessante para identificar esta classe, pois o modelo se mantém estável até um ponto de recall mais elevado. A estabilidade sugere que o modelo tem uma boa capacidade de generalizar para esta classe. De forma análoga à avaliação dessa classe na seção anterior, é relevante levar em consideração a quantidade de amostras reais no conjunto original, que era muito superior às classes minoritárias, que por sua vez, utilizaram amostras sintéticas.

Classe 1 (Secreto) - Demonstrada pela linha laranja, a classe 1 apresenta um desafio maior para a tarefa de classificação do modelo. Existe uma diminuição mais elevada da precisão ao tentar elevar o recall. É possível inferir que há uma maior ambiguidade ou sobreposição de características que identificam esta classe para o modelo, principalmente na

tarefa de diferenciar os dados como secretos ou ultrassecretos. A queda na precisão indica que o modelo pode apresentar mais falsos positivos ao tentar capturar informações secretas.

Classe 2 (Ultrassecreto) - A classe 2, em verde, apresenta desempenho similar à classe 1, mas ligeiramente superior. O modelo mostra uma queda leve e contínua na precisão à medida que o recall aumenta, mas parece manter uma precisão adequada em níveis elevados de recall, indicando uma melhor capacidade do modelo de identificar corretamente instâncias dessa classe sem sacrificar significativamente a precisão. Isso sugere que o classificador tem uma boa capacidade de identificar informações ultrassecretas, mas ainda há espaço para melhorias, especialmente na redução de falsos positivos.

4.4 Implicações dos resultados

Os resultados deste estudo mostram que o uso de aprendizado de máquina na classificação de informações sigilosas do Governo Federal pode trazer mais segurança e eficiência para a gestão desses dados. Com uma acurácia de 94,39%, o modelo Random Forest se mostrou eficaz na categorização automática de documentos. O uso de um modelo como este em sistemas governamentais levanta a possibilidade de evitar erros humanos e apoiar o cumprimento de regulações como a LAI e a LGPD de forma automática.

O modelo teve um desempenho satisfatório na identificação de documentos ultrassecretos, o que reforça seu potencial para proteger dados críticos e padronizar a classificação de sigilo dentro dos órgãos públicos. No entanto, um fator relevante para essa performance foi o uso da técnica SMOTE para balancear as classes, gerando amostras sintéticas para as categorias minoritárias no conjunto de dados original. O uso de amostras sintéticas pode aumentar a probabilidade de overfitting, onde o modelo aprende padrões das amostras artificiais e não generaliza bem para dados reais.

Desse modo, é especialmente importante considerar fatores como diversidade na coleta de dados e padronização, especificamente na representatividade das classes minoritárias, balanceamento dos grupos e estudo de técnicas mais avançadas para classificação. Além disso, expandir a base de dados com mais exemplos reais de diferentes órgãos governamentais pode ajudar a tornar a classificação mais robusta, permitindo que a automação contribua de fato para a segurança e governança da informação pública.

5 TRABALHOS RELACIONADOS

A classificação de informações sigilosas por aprendizado de máquina já foi explorada por García-Pablos, Pérez e Cuadros (2020) utilizando modelos de linguagem pré-treinados, como BERT (Representações Codificadoras Bidirecionais de Transformadores) associado aos avanços do Processamento de Linguagem Natural (PLN). Este estudo considera particularmente a tarefa de classificação de informações sensíveis contidas em documentos que devem passar por um processo de anonimização. Para atingir esse objetivo, os sistemas podem utilizar uma combinação de abordagens baseadas em regras ou aprendizado de máquina, ou uma combinação de ambos.

O trabalho avaliou a aplicação de BERT na anonimização de dados clínicos espanhóis. Experimentos comparando um modelo BERT pré-treinado com outros métodos mostraram que ele supera alternativas sem necessidade de trabalhar características específicas do domínio, alcançando alto recall e robustez frente a dados escassos. Além disso, a abordagem baseada em BERT apresentou resultados competitivos no desafio MEDDOCAN 2019, ficando próxima do sistema vencedor. Estudos futuros podem explorar novos modelos multi-linguísticos e técnicas de ajuste fino para aprimorar ainda mais o desempenho.

No contexto governamental, garantir a privacidade de informações sigilosas é essencial para atender a normas como a Lei Geral de Proteção de Dados (LGPD) e regulamentações específicas sobre segurança da informação.

A robustez do modelo BERT em cenários de dados escassos sugere que ele pode ser utilizado em contextos governamentais onde há limitação de bases de dados anotadas, um desafio comum em setores regulados. As técnicas apresentadas podem ser aplicadas à classificação de informações sigilosas de dados de instituições públicas para automatizar a proteção de documentos sensíveis, reduzir erros humanos e melhorar a conformidade com normas de segurança e privacidade.

5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Neste trabalho, avaliamos um modelo de aprendizado de máquina para a tarefa de classificação de informações sigilosas com dados do Governo Federal. De início, os dados coletados apresentaram um desbalanceamento significativo de classes, fator que pode influenciar diretamente no treinamento e performance do modelo.

O classificador com o algoritmo Random Forest desempenhou bem para as três classes, apresentando indicadores bons para precisão e recall, métricas que avaliam a predição de valores verdadeiramente positivos.

Algumas limitações deste trabalho encontram-se na quantidade de informações disponíveis, na ausência de padrão na estrutura dos dados publicados e dimensionalidade restritiva de cada coleção de dados. Estes quesitos estão diretamente ligados à seleção de características e complexidade que pode ser aplicada ao modelo.

Em trabalhos futuros, é relevante que a etapa de coleta de dados considere bases de informações classificadas de outros órgãos ou instituições públicas. Esta abordagem pode melhorar consideravelmente a precisão do modelo e diminuir possíveis vieses intrínsecos aos dados.

Para melhorar o modelo, futuros trabalhos podem explorar técnicas de balanceamento de classes mais avançadas, a incorporação de dados adicionais ou a utilização de algoritmos mais sofisticados, como redes neurais profundas ou modelos baseados em transformers (e.g., BERT).

Dessa forma, um modelo de aprendizado de máquina para classificação de informações sigilosas do Governo Federal surge como uma alternativa possível de ser aplicada em sistemas reais no apoio do cumprimento de regulamentações como a LAI e a LGPD, reduzindo erros humanos na correta classificação e melhorando a eficiência de sistemas e processos públicos, ao buscar auxiliar na classificação de documentos e dados sensíveis conforme normas e dispositivos legais, reduzindo o risco de acesso não autorizado e potencializando a proteção da informação.

REFERÊNCIAS

ALIFERIS, Constanti; SIMON, Gyorgy. Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI. **Computers in health care (New York)**, p. 477–524, jan. 2024.

BERRAR, Daniel et al. **Cross-validation**. 2019. Disponível em: https://dberrar.github.io/papers/Berrar_EBCB_2nd_edition_Cross-validation_preprint.pdf. Acesso em 12 de fevereiro de 2025.

BORRA, S.; CIACCIO, A. Measuring the Prediction Error. A Comparison of Cross-Validation, Bootstrap and Covariance Penalty Methods. **Computational Statistics & Data Analysis**, v. 54, p. 2976-2989. 2010.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **Journal of Machine Learning Research**, v. 13, n. Feb, p. 281–305, 2012.

BRASIL. **Estratégia de Governo Digital 2023-2026**. [S.l.: s.n.], 2024.Disponível em: https://www.gov.br/governodigital/pt-br/estrategias-e-governanca-digital/estrategianacional. Acesso em 22 de outubro de 2024.

CASTRO, C. L. DE .; BRAGA, A. P.. Aprendizado supervisionado com conjuntos de dados desbalanceados. **Sba: Controle & Automação Sociedade Brasileira de Automatica**, v. 22, n. 5, p. 441–466, set. 2011.

CHAWLA, N. V.; BOWYER, K. W.; KEGELMEYER, P. W.. Smote: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, jun. 2002.

CRISTÓVAM, José Sérgio da Silva; HAHN, Tatiana Meinhart. Administração Pública Orientada Por Dados: Governo Aberto E Infraestrutura Nacional De Dados Abertos. **Revista de Direito Administrativo e Gestão Pública**, v. 6, n. 1, p. 1–24, ago 2020. Disponível em: https://www.indexlaw.org/index.php/rdagp/article/view/6388. Acesso em 23 de outubro de 2024.

CRISTÓVAM, José Sérgio da Silva; SAIKALI, Lucas Bossoni; SOUSA, Thanderson Pereira de. **Governo Digital na Implementação de Serviços Públicos para a Concretização de Direitos Sociais no Brasil.** Sequência (Florianópolis). Programa de Pós-Graduação em Direito da Universidade Federal de Santa Catarina, n. 84, p. 209–242, jan. 2020. Disponível em: https://doi.org/10.5007/2177-7055.2020v43n89p209. Acesso em 21 de outubro de 2024.

DOSHI-VELEZ, Finale; KIM, Been. **Towards A Rigorous Science of Interpretable Machine Learning**. [S.l.: s.n.], 2017. Disponível em: https://arxiv.org/abs/1702.08608. Acesso em 12 de novembro de 2024.

ERICKSON, Bradley J.; KITAMURA, Felipe. Magician’s Corner: 9. Performance Metrics for

Machine Learning Models. **Radiology: Artificial Intelligence**, v. 3, n. 3, 2021. Disponível em: https://doi.org/10.1148/ryai.2021200126. Acesso em 11 de novembro de 2024.

FARQUAD, M.A.H.; BOSE, Indranil. Preprocessing unbalanced data using support vector machine. **Decision Support Systems**, v. 53, n. 1, p. 226–233, 2012. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167923612000425. Acesso em 11 de novembro de 2024.

GARCÍA-PABLOS, Aitor; PÉREZ, Naiara; CUADROS, Montse. **Sensitive Data Detection and Classification in Spanish Clinical Text:** Experiments with BERT. 2020. Disponível em: https://api.semanticscholar.org/CorpusID:212628622. Acesso em 9 de janeiro de 2025.

GOUVEIA, Luís Borges. **Local e-Government-A governação digital na autarquia**. SPI/Principia, 2004. Disponível em: https://bdigital.ufp.pt/handle/10284/263. Acesso em 11 de novembro de 2024.

HAYKIN, S. **Redes neurais:** princípios e prática. [S.l.]: Bookman Editora, 2007.

HE, H; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, *v.* **21**, n. 9, p. 1263–1284. 2009.

HULJANAH, Mia et al. Feature Selection using Random Forest Classifier for Predicting Prostate Cancer. **IOP Publishing,** v. 546, n. 5, p. 052031, 2019. Disponível em: https://dx.doi.org/10.1088/1757-899X/546/5/052031. Acesso em 10 de novembro de 2024.

KOHAVI, Ron. **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. Mar. 1995. Disponível em: https://www.researchgate.net/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection. Acesso em 19 fev. de 2025.

KIM, Ji-Hyun. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. **Computational Statistics & Data Analysis**, v. 53, n. 11, p. 3735-3745. 2009.

LIMA, Tiago et al. Previsão de óbito e importância de características clínicas em idosos com COVID19 utilizando o Algoritmo Random Forest. **Revista Brasileira de Saúde Materno Infantil**. v. 21, p 445-451, mar. 2021. Disponível em: http://higia.imip.org.br/handle/123456789/878. Acesso em 15 nov. 2024.

LUDERMIR, Teresa Bernarda. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. **Estudos Avançados**, Instituto de Estudos Avançados da Universidade de São Paulo, v. 35, n. 101, p. 85–94, jan. 2021. Disponível em: https://doi.org/10.1590/s0103-4014.2021.35101.007. Acesso em 15 nov. 2024.

MIAO, Jiaju; ZHU, Wei. Precision–recall curve (PRC) classification trees. **Evolutionary Intelligence**, v. 15, p. 1545–1569, 2022. Disponível em: https://arxiv.org/abs/2011.07640. Acesso em 13 de fevereiro de 2025. Acesso em 15 nov. 2024.

MICROSOFT. **Treinar um modelo de machine learning usando validação cruzada**. Disponível em: https://learn.microsoft.com/pt-br/dotnet/machine-learning/how-to-guides/train-machine-learning-model-cross-validation-ml-net. Acesso em 15 nov. 2024.

MINISTÉRIO DA FAZENDA. **Rol de informações Classificadas e Desclassificadas**. Disponível em: https://www.gov.br/fazenda/pt-br/acesso-a-informacao/informacoes-classificadas. Acesso em 27 de fev. 2025.

MINISTÉRIO DA JUSTIÇA. **Rol de informações Classificadas e Desclassificadas**. Disponível em: https://dados.mj.gov.br/dataset/rol-de-informacoes-classificadas-de-desclassificadas. Acesso em 27 de fev. 2025.

MINISTÉRIO DA SÁUDE. **Rol de informações Classificadas e Desclassificadas**. Disponível em: https://www.gov.br/saude/pt-br/acesso-a-informacao/informacoes-classificadas/rol-de-informacoes-classificadas. Acesso em 27 de fev. 2025.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003. Disponível em: https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf. Acesso em: 13 de fevereiro de 2025.

PATTERSON, Josh; GIBSON, Adam. **Deep Learning: A Practitioner's Approach**. [S.l.]: "O'Reilly Media, Inc.", 2017.

ROSA, Ana Paula Schneid Afonso da et al. Uso de técnicas de aprendizado de máquina para classificação de fatores que influenciam a ocorrência de dermatites ocupacionais. **Revista Brasileira de Saúde Ocupacional**, 2023. Disponível em: https://api.semanticscholar.org/CorpusID:258409986. Acesso em: 13 de fevereiro de 2025.

REIS, Carlos Henrique. **Otimização de Hiperparâmetros em Redes Neurais Profundas**. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Centro de Ciências Exatas e Naturais, Universidade Federal de Itajubá- UNIFEI, Itajubá-MS. 2018. 72 f.

VIANA, Ana Cristina Aguilar. Transformação digital na administração pública: do governo eletrônico ao governo digital. **Revista Eurolatinoamericana de Derecho Administrativo**, 2021. Disponível em: https://api.semanticscholar.org/CorpusID:246723112. Acesso em: 13 de fevereiro de 2025.