Identificação Automática de Registros Espúrios no Cadastro Ambiental Rural via Aprendizagem de Máquina

Fernando E. M. Borges * **Danton D. Ferreira** **
Antônio C. S. C. Junior *** **Ricardo R. Magalhães** ****

* *Programa de Pós-Graduação em Engenharia Agrícola, Universidade Federal de Lavras, Minas Gerais, Brasil, (e-mail: fernandoelias.mb@gmail.com).*
** *Departamento de Automática, Universidade Federal de Lavras, Lavras, Minas Gerais, Brasil, (e-mail: danton@ufla.br)*
*** *Agência Zetta de Inovação, Universidade Federal de Lavras, Minas Gerais, Brasil, (e-mail: antoniocoutojr.ti@fundecc.org.br)*
**** *Departamento de Automática, Universidade Federal de Lavras, Lavras, Minas Gerais, Brasil, (e-mail: ricardorm@ufla.br)*

Abstract: The Rural Environmental Registry (CAR) is a mandatory electronic registry for all rural properties in Brazil. Currently, the CAR is gaining importance in land use monitoring and environmental regularization. However, the regularization analyses of the registry are done manually, requiring a significant amount of time and skilled labor for this task. This study aims to use machine learning algorithms for the classification and identification of erroneously filled (spurious) records. As a database, real CAR records previously manually labeled were used. Subsequently, six classification models were applied, and the generated results were evaluated and compared with each other. The classification models, in general, achieved evaluation metrics above 80%, and the Gradient Boosting obtained all metrics above 90% for the test dataset. The results obtained through the models demonstrated the feasibility of implementing the model in a real CAR scenario, through a REST API. Thus, creating the machine learning module in an API, it is possible to develop an application to agilize and improve the processes in the Rural Environment Registry in Brazil.

Resumo: O Cadastro Ambiental Rural (CAR) consiste em um registro eletrônico obrigatório para todos os imóveis rurais no Brasil. Atualmente, o CAR vem ganhando importância no monitoramento do uso de terra e regularização ambiental. Todavia, as análises de regularização do cadastro são feitas de forma manual, demandando elevado tempo de análise e mão-de-obra qualificada para esta tarefa. Este trabalho tem por objetivo o uso de algoritmos de aprendizagem de máquina para a classificação e identificação dos registros preenchidos erroneamente (espúrios). Como base de dados, foram utilizados registros reais do CAR previamente rotulados manualmente. Em seguida, foram aplicados 6 modelos de classificação e os resultados gerados foram avaliados e comparados entre si. Os modelos de classificação, em geral, atingiram métricas de avaliação acima de 80% e o *Gradient Boosting* obteve todas as métricas acima de 90% para o conjunto de teste. Os resultados obtidos por meio dos modelos apresentaram viabilidade de implementação do modelo em um cenário real do CAR, por meio de API REST. Assim, com a criação do módulo de aprendizagem de máquina em uma API, é possível desenvolver uma aplicação para agilizar e melhorar os processos do Cadastro Ambiental Rural no Brasil.

Keywords: Rural Environmental Registry; Machine Learning; Data Science; Data Classification; Data Mining.

Palavras-chaves: Cadastro Ambiental Rural; Aprendizagem de Máquina; Ciência de dados; Classificação de Dados; Mineração de Dados.

1. INTRODUÇÃO

Criado com o objetivo de monitorar o uso de terra nas propriedades rurais, auxiliar no combate ao desmatamento ilegal e invasões de terra, em 2012, foi criado o Cadastro Ambiental Rural (CAR), durante a promulgação do Código Florestal (Brasil, 2012). O CAR consiste em um registro eletrônico público e obrigatório para todos os territórios rurais do Brasil, onde os proprietários destes imóveis precisam realizar o cadastro via sistema eletrônico do CAR, o SiCAR. Por meio do SiCAR, o cadastrante (este podendo ser o proprietário ou outra pessoa contratada por este para realizar o cadastro) insere dados referentes ao imóvel rural, uso de terra, dados de geolocalização da propriedade, dentre outras informações pertinentes. Além do registro dos imóveis rurais, os dados do CAR, uma vez sendo de domínio público, podem ser utilizados para estudos de monitoramento do uso de terra no Brasil. Em um destes estudos, L'Roe et al. (2016) estudaram a influência da implementação do CAR em ações contra desmatamento e outras ilegalidades. Outros estudos reforçam o monitoramento geográfico das áreas ocupadas com o CAR, visando uma melhoria do uso de terra, a tornando mais sustentável, além do combate ao desmatamento (Arvor et al., 2021; dos Santos et al., 2020)

O monitoramento geográfico das áreas é de suma importância para o entendimento das implicações da implementação do CAR. Além disso, validar os cadastros realizados também é de suma importância para garantir a regularização dos imóveis rurais no país. De acordo com o boletim informativo do CAR do mês de abril de 2023 (Serviço Florestal Brasileiro, 2023), dos quase 7 milhões de cadastros, 26% destes passaram por algum tipo de análise e menos de 1% concluíram a regularidade ambiental. Tais informações podem ser observadas no infográfico disposto na Figura 1.

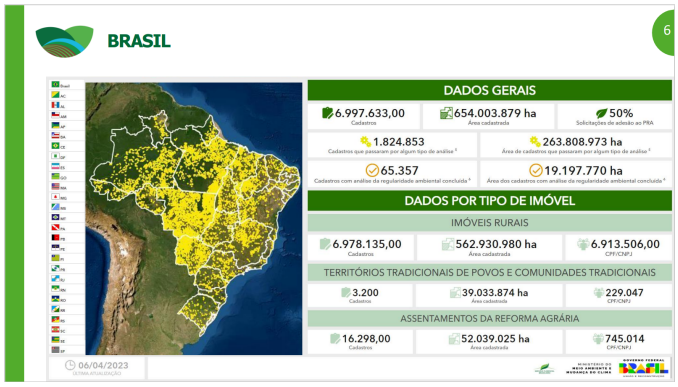


Figura 1. Infográfico com os números do CAR. Fonte: Serviço Florestal Brasileiro (2023)

Além dos dados supracitados, cabe salientar que o Brasil possui mais de 8 milhões de imóveis rurais, de acordo com dados da Receita Federal (Receita Federal, 2022). Portanto, elaborar um sistema que agilize o processo de análise dos registros do CAR é crucial para a melhoria da plataforma.

Analisar os registros do CAR de forma manual requer pessoal qualificado para determinada tarefa, o qual precisará identificar os padrões que qualificam determinado cadastro a ser aprovado ou não. Associado o desafio de qualificação de pessoal ao imenso número de registros, trata-se de uma tarefa que demandará muito tempo e custo para sua conclusão. Dada esta problemática, este trabalho propõe o uso de algoritmos de aprendizagem de máquina para identificação de registros espúrios do CAR. O objetivo deste trabalho consiste em unir técnicas de ciência de dados e aprendizagem de máquina para identificar quais registros do CAR serão aprovados ou cancelados, por meio de modelos de classificação de dados. A principal contribuição deste trabalho é o uso de algoritmos de classificação aos cadastros do CAR, algo, até o momento, sem registro de trabalhos realizados na literatura.

A aplicação de aprendizagem de máquina para classificação de registro em formato tabular é uma das mais utilizadas na literatura, para as mais variadas finalidades. Na área da saúde, pode ser mencionado o trabalho reportado por Schneider et al. (2021), onde os autores aplicaram modelos como Máquinas de Vetor Suporte, *Random Forest*, *XGBoost* e modelos de Redes Neurais profundas a fim de classificar doenças inflamatórias intestinais pediátricas, obtendo valores de precisão de 90%.

Outra aplicabilidade pode ser vista na área de segurança de informação, como no trabalho desenvolvido por Syeda and Asghar (2024), no qual foram aplicados modelos como *Random Forest*, *XGBoost*, k-vizinhos mais próximos, Regressão Logística, dentre outros para a identificação de *malware* em chamadas de APIs , atingindo taxas de verdadeiro positivo de 93% e falso positivo de menos de 1%.

Os trabalhos mencionados podem, de maneira sucinta, apresentar a boa aplicabilidade dos modelos de aprendizagem de máquina em cenários variados, como auxílio na tomada de decisão de identificação de registros anômalos.

O presente trabalho segue dividido em 3 seções posteriores: A seção 2 apresenta os materiais e métodos aplicados para o trabalho, na seção 3 estão dispostos os resultados e discussões, enquanto as conclusões e perspectivas para o trabalho estão descritos na seção 4.

2. MATERIAL E MÉTODOS

Para a realização deste trabalho, todo o procedimento desde a padronização dos dados, pré-processamento, implementação dos modelos de aprendizagem e visualização dos resultados foi feito utilizando a linguagem Python, versão 3.12. Como bibliotecas utilizadas, destacam-se: o Numpy para cálculos algébricos e matriciais; Pandas para análise de dados, uso de dataframes e visualização de dados; Scikit-Learn para modelagem dos modelos de classificação; Matplotlib e Seaborn para geração de gráficos.

2.1 Base de Dados e Pré-Processamento

A base dados utilizada consiste em registros reais fornecidos pela Agência Zetta de Inovação da Universidade Fede-

ral de Lavras (UFLA), contendo registros do CAR de 19 estados brasileiros, contabilizando 32.458 registros no total. Como atributos existentes na base de dados, estão dispostos atributos referentes à área dos imóveis rurais, vértices do polígono do desenho do terreno inserido, questionário acerca da regularidade da propriedade, feições do terreno (rios, tipo de vegetação, etc.), retificações realizadas pelo cadastrante no SiCAR [1] , além de sobreposições com outros imóveis rurais, feito por meio de cruzamento dos dados de georreferenciamento. O número de registros corresponde a um baixo percentual em relação ao número de registros totais do CAR devido à necessidade dos registros estarem devidamente rotulados para o treinamento dos classificadores. Tal cenário indica uma limitação na aquisição dos dados, além de uma motivação para a implementação da aplicação.

Antes da execução dos procedimentos de pré-processamento, os dados foram divididos em treinamento (onde serão realizados os processos exploratórios, de limpeza e padronização dos dados, assim como o treinamento com validação cruzada dos classificadores) e teste (dados novos que são apresentados aos modelos já treinados). A escolha dos dados de treinamento se deu por amostragem aleatória simples, sem reposição, estratificada por classe. A proporção se deu em 80% dos registros do CAR para treinamento e os 20% restantes para teste. Tal divisão foi realizada em uma única repetição de forma a garantir que os dados de teste não tenham sido vistos anteriormente pelos classificadores.

De maneira a padronizar o conjunto de dados, foi necessária a realização de alguns procedimentos iniciais. Primeiramente, foi realizada a remoção de atributos que não continha informações pertinentes para a classificação como a 'UF do imóvel', 'município', 'id do imóvel', além da remoção de informações sensíveis, como o nome do cadastrante.

Após a filtragem inicial, foi realizada a codificação dos atributos do questionário, onde os valores de 'sim', 'não' e 'não informar' foram codificados para 2, 1 e 0, respectivamente. Outra adaptação foi realizada nos atributos referentes à área do imóvel rural, uma vez que houveram mais de uma área registrada, assim, foi atribuída como a área final, o somatório das áreas inseridas. Por fim, foi codificada a variável de saída, denominada 'condicao', onde foi atribuído 0 para os registros cancelados e 1 para os registros aprovados. O número de registros por estado na base de dados segue disposto na Tabela 1

Após os procedimentos de pré-processamento, foram obtidos 83 atributos de entrada. Em seguida, foi aplicado o método da Razão de Discriminação de Fisher (FDR - *Fisher's Discriminant Ratio*) (Duda and Hart, 2001) para a seleção dos atributos mais relevantes para a classificação. A seleção se deu da seguinte forma: primeiramente, foi calculado o vetor de pesos do FDR, em seguida, os pesos foram ordenados de forma decrescente e foram selecionados os atributos em intervalos subdivididos: primeiramente de 10 em 10 (ex.: 10, 20...), depois foi coletado o intervalo de melhor resultado (ex.: 20 a 30 atributos). Depois da escolha do melhor intervalo, o intervalo selecionado foi subdividido em intervalos de 5 em 5 e depois variado unitariamente. Após a seleção de atributos, foi realizada a transformação logarítmica nos atributos contínuos, dado que estes possuíam elevada assimetria.

[1] https://www.car.gov.br/#/

Tabela 1. Números da base de dados por estado.

UF	Aprovados	Cancelados	Total
AC	14	322	336
AL	0	100	100
AP	0	258	258
CE	1456	2535	3991
ES	0	540	540
GO	40	578	618
MA	20	781	801
PA	159	3819	3978
PB	51	280	331
PE	0	69	69
PI	0	162	162
RJ	160	400	560
RN	0	138	138
RO	323	1571	1894
RR	0	19	19
SE	2	87	89
SP	632	4983	5615
TO	3454	9505	12959
Total	6311	26147	32458

2.2 Modelos de Aprendizado

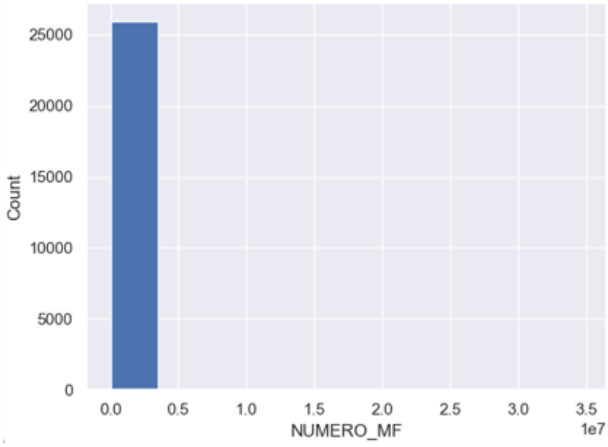
Como modelos de aprendizagem, foram aplicados 6 algoritmos para a realização de testes comparativos entre estes. A fim de elencar o modelo, entre parênteses estará uma sigla de 3 letras rotulando o modelo para as análises posteriores. Os modelos utilizados foram: *AdaBoost* (ABC) (Hastie et al., 2009); *Gradient Boosting* (GBT) (Friedman, 2002); Regressão logística (LRC) (Evsukoff, 2020); Redes Neurais do tipo *Perceptron* Multicamadas (MLP) (Haykin, 2007); *Random Forests* (RFC) (Breiman, 2001); e o classificador de Máquinas de Vetores de Suporte (SVM) (Vapnik, 1998). A ideia foi elencar diversos tipos de modelos, sejam lineares, não lineares e máquinas de comitê (*ensembles*), dada a capacidade que cada modelo tem de performar melhor em certo tipo de base de dados. Uma vez que a base de dados do CAR consiste em dados discretos e contínuos, testar diferentes tipos de algoritmos é importante para a realização dos testes.

2.3 Avaliação dos Classificadores

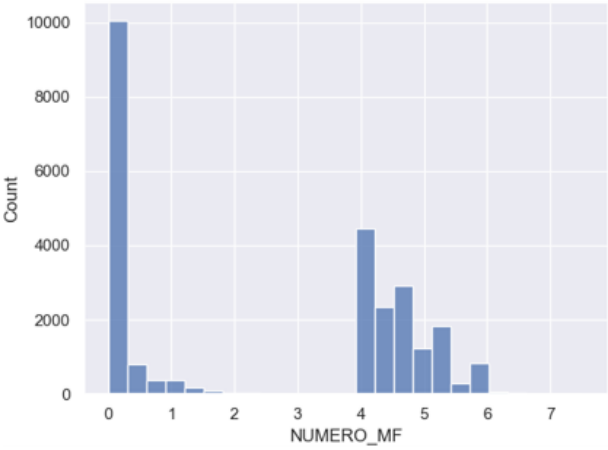
O ajuste dos hiperparâmetros dos modelos, ou seja, os parâmetros que devem ser imputados pelo usuário foi realizado manualmente, via tentativa e erro, buscando obter o modelo de melhor acurácia. Em seguida foi realizado o ajuste dos parâmetros dos classificadores, ou seja, os pesos do modelo de aprendizagem. Para este ajuste, foi feito o uso de validação cruzada do tipo *k-fold* com 10 *folds*, onde o critério de escolha do melhor modelo se desse pelo *fold* de maior acurácia. Foram levantados os resultados médio e o desvio-padrão para a validação cruzada, utilizando o *fold* de teste e os resultados obtidos no conjunto de teste do modelo. Como métricas de avaliação foram utilizadas a acurácia média do modelo (ACC), Precisão e Recall para cada classe, F1-Score, índice *Kappa* e a área sob a curva ROC (AUC).

3. RESULTADOS E DISCUSSÕES

Com relação aos resultados de pré-processamento, constatou-se a importância dos procedimentos de limpeza, codificação e transformação dos atributos. De maneira a ilustrar o efeito da transformação de atributos, sobretudo para os casos de assimetria, foram gerados histogramas comparando a distribuição das variáveis antes e depois da transformação logarítmica. Os comparativos antes e depois da transformação para as variáveis 'NUMERO MF', 'AREA DOC', 'ARL TOTAL' estão dispostos, respectivamente, nas Figuras 2, 3 e 4. A aplicação da transformação logarítmica reduziu as discrepâncias entre as escalas de cada atributo, o que pode melhorar o desempenho do modelo, uma vez que todos os atributos terão, aproximadamente, a mesma relevância para o modelo.



(a) Variável sem transformação

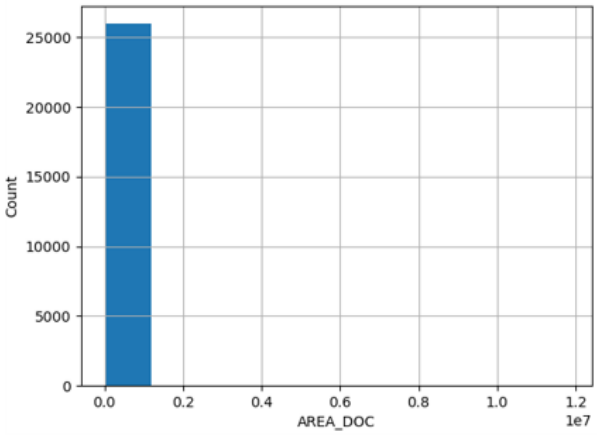


(b) Variável transformada

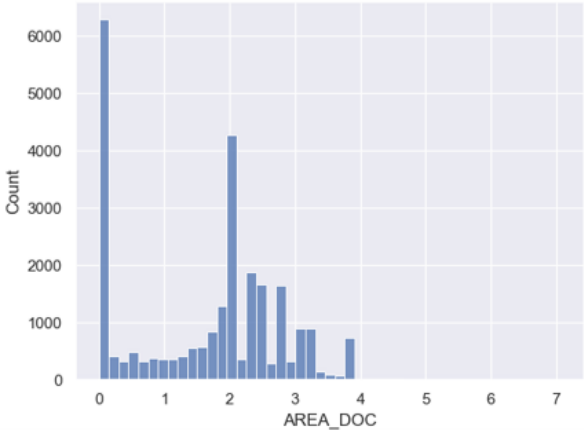
Figura 2. Efeito da transformação logarítmica para a variável 'NUMERO MF'

Seguido a transformação dos dados, foi realizada a divisão dos dados em conjuntos de treinamento + validação cruzada e teste. A fim de ilustrar os números de cada conjunto, foi gerada uma tabela contendo os tamanhos de cada conjunto. Os números dos conjuntos de treinamento e teste para cada classe podem ser vistos na Tabela 2.

Em posse dos conjuntos de treinamento, foi realizado o ajuste manual dos hiperparâmetros. Os valores de cada



(a) Variável sem transformação



(b) Variável transformada

Figura 3. Efeito da transformação logarítmica para a variável 'AREA DOC'

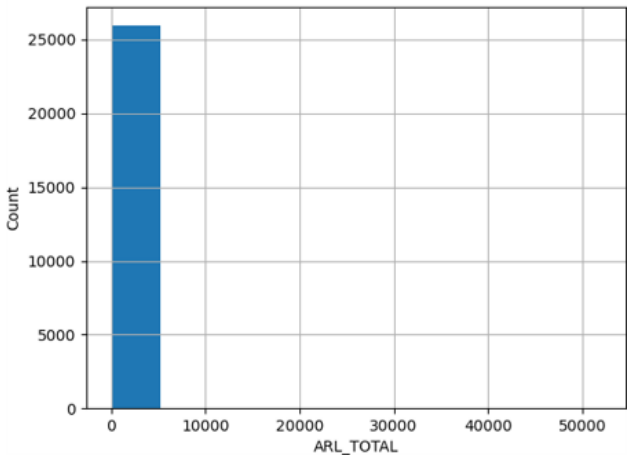
Tabela 2. Conjuntos de treinamento e teste utilizados

Classe	Treinamento + Validação	Teste	Total
Aprovados – 1	50418	1263	6311
Cancelados – 0	20917	5230	26147
Total	25965	6493	32458

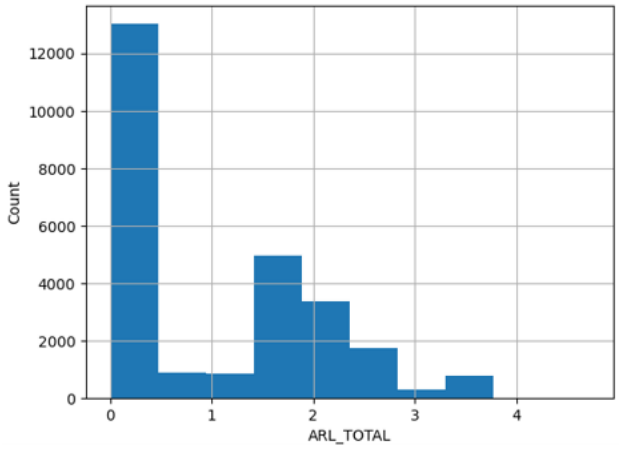
hiperparâmetro ajustado para cada classificador estão dispostos na Tabela 3.

onde o otimizador 'lbfgs' se trata de um modelo de ajuste de parâmetros de segunda ordem da família de métodos quasi-Newton; o parâmetro *scale*, quando utlizado, aplica a seguinte formulação para obtenção do hiperparâmetro: $\gamma = 1/(n_{atributos}\sigma^2(\mathbf{X}))$, no qual $\sigma^2(\mathbf{X})$ é a variância de todo o conjunto de dados.

Após o ajuste dos hiperparâmetros, foram realizados os procedimentos de treinamento, validação e teste dos modelos, cabe salientar que o conjunto de teste é composto de registros novos ao classificador, ou seja, que em nenhum momento tais registros foram utilizados no ajuste de parâmetros enquanto o conjunto de validação é composto pelos registros de teste durante a validação cruzada *k-fold*. Os resultados de validação estão dispostos no formato média ± desvio-padrão. Na Tabela 5 estão dispostos os resultados percentuais para cada um dos modelos utilizados. Os re-



(a) Variável sem transformação



(b) Variável transformada

Figura 4. Efeito da transformação logarítmica para a variável 'ARL TOTAL'

Tabela 4. Hiperparâmetros dos modelos utilizados

Modelo	Hiperparâmetro	Valor
ABC	Medida de avaliação	Índice gini
	Divisão de folhas	Melhor valor
	Profundidade máxima	9
	Número de árvores	300
GBT	Profundidade máxima	5
	Taxa de aprendizado	0,2
	Número de árvores	300
LRC	C	1438,45
	Otimizador	Lbfgs
	Regularização	l2 (*ridge*)
MLP	Neurônios na camada oculta	25
	Função de ativação	tangente hiperbólica
	Otimizador	Lbfgs
	Taxa de aprendizado	Adaptativo
RFC	Profundidade máxima	19
	Medida de avaliação	Entropia
	Número de árvores	100
SVM	C	1000
	γ	*Scale*

sultados apresentados, em geral, foram satisfatórios, onde todos os modelos, exceto o LRC obtiveram mais de 80% no índice *Kappa*, o que indica um classificador de ótimo desempenho (Landis and Koch, 1977). Outro ponto, foi o efeito positivo da classificação logarítmica que atenuou os efeitos de assimetria, acarretando em melhoria no desempenho. A fim de apontar um classificador de melhor desempenho geral, dando ênfase ao conjunto de teste e à métrica do índice *Kappa*, o modelo GBT foi o que melhor performou, com todas as medidas de avaliação acima dos 90%. A fim de ilustrar a curva ROC de cada modelo, foi gerada a curva ROC dos classificadores para comparação de seu comportamento em função dos verdadeiros e falsos positivos. Na Figura 5 estão dispostas as curvas ROC de todos os classificadores utilizados.

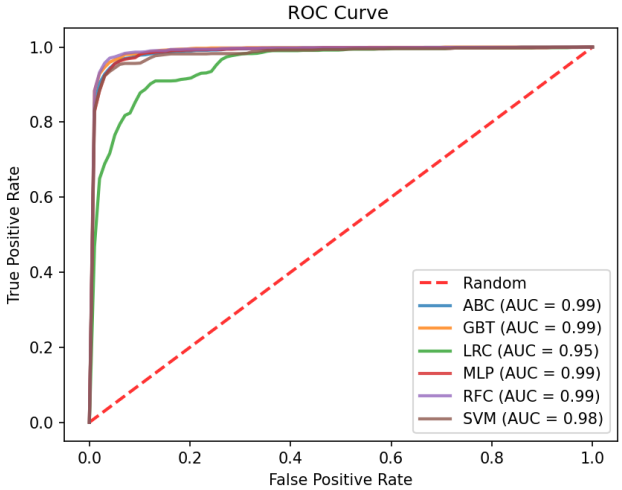


Figura 5. Curvas ROC obtidas para os classificadores considerando o conjunto de teste.

Como pontos de destaque, o conjunto de operações realizadas, tanto de pré-processamento, processamento e pós-processamento, indicou que uma padronização das ferramentas de ciência de dados e aprendizagem de máquina, pode ser viável para os registros do CAR, uma vez que, utilizando a base pública de dados com alguns refinamentos, conforme fornecido neste trabalho, atingiu resultados acima de 90%. Tal situação demonstra uma viabilidade do uso do sistema em uma aplicação real, todavia, deve-se ater no tocante aos falsos positivos do modelo, ou seja, classificar um registro como aprovado, sendo o mesmo com erros. Desta forma, mapear um ponto de operação do modelo GBT e sua saída, se faz importante para a operação em um cenário real.

Como limitação prática da aplicação, cabe mencionar o ajuste dos hiperparâmetros, os quais, no momento do estudo, foram feitos via tentativa e erro para otimizar o custo computacional. Todavia, o uso de algoritmos de busca em servidores robustos pode melhorar o processo de treinamento do modelo em trabalhos futuros.

4. CONCLUSÕES E PERSPECTIVAS

Este trabalho tem como objetivo principal a implementação de modelos de classificação na identificação de registros a serem aprovados e cancelados dentro do Cadastro Ambiental Rural. Foram aplicadas diversas técnicas englobando Ciência de Dados e Aprendizagem de Máquina e foram gerados 6 classificadores. Ao final do experimento, durante

Tabela 5. Resultados de classificação para os conjuntos de treinamento (sob validação cruzada) e teste

Conjunto de dados	Modelo	ACC	AUC	F1-Score	*Kappa*	Precisão: Classe 0	Precisão: Classe 1	Recall: Classe 0	Recall: Classe 1
Validação Cruzada	ABC	96,3 ± 0,2	99,0 ± 0,1	97,8 ± 0,1	88,2 ± 0,6	97,0 ± 0,2	93,7 ± 0,8	98,6 ± 0,2	87,2 ± 0,7
	GBT	96,7 ± 0,3	99,3 ± 0,1	**98,0 ± 0,2**	89,4 ± 1,1	**97,5 ± 0,2**	93,2 ± 1,2	98,4 ± 0,3	**89,7 ± 1,0**
	LRC	91,3 ± 0,6	95,4 ± 0,5	94,7 ± 0,4	71,3 ± 2,0	93,6 ± 0,5	80,5 ± 2,3	95,7 ± 0,6	73,1 ± 2,2
	MLP	96,1 ± 0,4	99,0 ± 0,2	97,6 ± 0,3	87,2 ± 1,5	97,1 ± 0,5	91,6 ± 1,0	98,0 ± 0,3	87,9 ± 2,1
	RFC	**96,8 ± 0,2**	**99,4 ± 0,0**	**98,0 ± 0,1**	**89,4 ± 0,8**	97,3 ± 0,3	**94,5 ± 0,9**	**98,8 ± 0,2**	88,5 ± 1,1
	SVM	95,9 ± 0,3	98,1 ± 0,2	97,5 ± 0,2	86,7 ± 0,9	96,8 ± 0,2	92,0 ± 1,3	98,2 ± 0,3	86,5 ± 1,0
Teste	ABC	96,6	98,9	97,9	89,1	97,3	93,7	98,5	88,7
	GBT	**97,0**	**99,2**	**98,1**	**90,5**	**97,8**	93,7	98,5	**91,0**
	LRC	91,2	95,3	94,6	71,1	93,6	80,3	95,6	73,0
	MLP	96,1	98,9	97,5	87,3	97,1	91,6	98,0	87,9
	RFC	96,9	**99,2**	**98,1**	89,9	97,3	**94,7**	**98,8**	89,0
	SVM	96,2	95,2	97,6	87,8	97,0	92,6	98,3	87,7

a análise dos resultados, pode-se observar que o modelo *Gradient Boosting* obteve resultados acima de 90%, o que aponta uma viabilidade do seu uso em uma aplicação real. Cabe salientar que o experimento utilizou 100% de dados reais, sem o uso de qualquer simulação ou geração de dados sintéticos.

Para projetos futuros, tem-se por objetivo incrementar este sistema em um modelo de arquitetura de microsserviços, via uma API REST. Além disso, se faz importante a aplicação de melhores ajustes nos hiperparâmetros, como o uso de algoritmos evolucionários, por exemplo. Outro ponto de destaque para os próximos passos é o ajuste do limiar de classificação ponderado de acordo com o custo de cada decisão, minimizando o custo das decisões incorretas. Tais ajustes e implementações beneficiarão o Brasil, por meio de um melhor monitoramento de terra e validação dos registos do CAR, além dos proprietários dos imóveis rurais que terão agilidade em seu processo de regularização ambiental.

AGRADECIMENTOS

Os autores gostariam de agradecer à Universidade Federal de Lavras, Agência Zetta de Inovação, FAPEMIG, CAPES e ao CNPq por apoiarem este trabalho.

REFERÊNCIAS

Arvor, D., Silgueiro, V., Nunes, G.M., Nabucet, J., and Dias, A.P. (2021). The 2008 map of consolidated rural areas in the brazilian legal amazon state of mato grosso: Accuracy assessment and implications for the environmental regularization of rural properties. *Land Use Policy*, 103, 105281.

Brasil (2012). Lei nº 12.651, de 25 de maio de 2012. *Diário Oficial da República Federativa do Brasil*. URL `http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2012/Lei/L12651.htm`.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

dos Santos, P.P., de Jesus Júnior, W.C., de Almeida Telles, L.A., de Souza, M.H., da Silva, S.F., dos Santos, A.R., et al. (2020). Geotechnologies applied to analysis of the rural environmental cadastre. *Land Use Policy*, 105127.

Duda, R.O. and Hart, P.E. (2001). *Pattern Classification*. John Wiely and Sons, New York, USA, 2 edition.

Evsukoff, A.G. (2020). *Inteligência Computacional—Fundamentos e Aplicações*. E-Papers, Rio de Janeiro, RJ, Brazil.

Friedman, J.H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4), 367–378.

Hastie, T., Tibshirani, R., Friedman, J.H., and Friedman, J.H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, second edition.

Haykin, S. (2007). *Redes neurais: princípios e prática*. Bookman Editora, Porto Alegre.

Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

L'Roe, J., Rausch, L., Munger, J., and Gibbs, H.K. (2016). Mapping properties to monitor forests: Landholder response to a large environmental registration program in the brazilian amazon. *Land Use Policy*, 57, 193–203.

Receita Federal (2022). Dados estatisticos do cafir. URL `https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/cadastros/portal-cnir/estatisticas-e-dados-abertos/dados-estatisticos-do-cafir`.

Schneider, N., Sohrabi, K., Schneider, H., Zimmer, K.P., Fischer, P., De Laffolie, J., and Group, C.G.S. (2021). machine learning classification of inflammatory bowel disease in children based on a large real-world pediatric cohort cedata-gpge® registry. *Frontiers in Medicine*, 8, 666190.

Serviço Florestal Brasileiro (2023). Numeros do cadastro ambiental rural. URL `https://www.florestal.gov.br/numeros-do-car`.

Syeda, D.Z. and Asghar, M.N. (2024). Dynamic malware classification and api categorisation of windows portable executable files using machine learning. *Applied Sciences*, 14(3), 1015.

Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley, New York.