FERNANDO ELIAS DE MELO BORGES

APLICAÇÃO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA IDENTIFICAÇÃO DE REGISTROS ESPÚRIOS NO CADASTRO AMBIENTAL RURAL

LAVRAS – MG

2022

FERNANDO ELIAS DE MELO BORGES

APLICAÇÃO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA IDENTIFICAÇÃO DE REGISTROS ESPÚRIOS NO CADASTRO AMBIENTAL RURAL

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas e Automação da Universidade Federal de Lavras como parte das exigências para a obtenção do título de Mestre.

Prof. Dr. Danton Diego Ferreira

Orientador

LAVRAS – MG

2022

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Borges, Fernando Elias de Melo
Aplicação de algoritmos de aprendizagem de máquina na identificação de registros espúrios no Cadastro Ambiental Rural / Fernando Elias de Melo Borges. 2022.
90 p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2022.
Orientador: Prof. Dr. Danton Diego Ferreira.
Bibliografia.

1. Cadastro Ambiental Rural. 2. Ciência de dados. 3. Aprendizagem de Máquina Interpretável. I. Ferreira, Danton Diego. II. Título.

FERNANDO ELIAS DE MELO BORGES

APLICAÇÃO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA IDENTIFICAÇÃO DE REGISTROS ESPÚRIOS NO CADASTRO AMBIENTAL RURAL
APPLICATION OF MACHINE LEARNING ALGORITHMS TO IDENTIFY SPURIOUS RECORDS IN THE RURAL ENVIRONMENTAL REGISTRY

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas e Automação da Universidade Federal de Lavras como parte das exigências para a obtenção do título de Mestre.

APROVADA em 29 de Julho de 2022.

Prof. DSc. Danton Diego Ferreira	UFLA
Prof. DSc. Alexandre Gonçalves Evsukoff	UFRJ
Prof. DSc. Wilian Soares Lacerda	UFLA

Prof. Dr. Danton Diego Ferreira
Orientador

LAVRAS – MG
2022

Dedico esta dissertação aos meus pais, Fernando Antônio e Cleide e à minha irmã Letycia, por todo o apoio dado desde sempre.

AGRADECIMENTOS

Estou aqui concluindo mais uma etapa em um período tão conturbado, globalmente e localmente para mim, passando por diversas situações, mas creio que estou conseguindo sair melhor do que comecei, mais maduro e preparado para o que poderá vir.

Durante o mestrado, pessoas importantes passaram de maneira a contribuir para este trabalho e para mim mesmo como pessoa e gostaria, aqui, de fazer meus singelos agradecimentos.

Primeiramente aos meus pais e minha irmã, por todo o carinho, paciência e apoio dados até aqui, por sempre me incentivarem a buscar os meus objetivos.

Agradeço aos docentes do PPGESISA que sempre estiveram dispostos a ajudar e contribuir durante minha jornada, agradeço, em especial, ao professor Danton Diego Ferreira por toda a parceria ao longo destes e anos, pela confiança depositada e por acreditar no meu potencial.

Agradeço ao pessoal da Agência Zetta de Inovação pelo aporte à esta pesquisa e, sobretudo ao Antônio Couto Júnior e ao professor Erick Maziero pelas contribuições dadas à este trabalho. Agradeço, também, às contribuições dadas pelo professor Alexandre Evsukoff para a melhoria deste trabalho.

Aos Companheiros da República Vaca-H pelo apoio moral e pela paciência ao longo destes anos, obrigado por tudo! No mais, agradeço à todos e todas que contribuíram, seja profissionalmente ou pessoalmente e, caso eu tenha esquecido de alguém, fica meu pedido de desculpas (risos).

Muito obrigado!

“A evolução não consiste chegar até o fim, à perfeição, mas sim nos esforços contínuos em melhorar o que já existe.”

- Autor Desconhecido.

RESUMO

O Cadastro Ambiental Rural (CAR) consiste em um registro público eletrônico obrigatório para todos os imóveis rurais do território brasileiro, integra informações ambientais das propriedades, auxiliando no monitoramento ambiental e contribui em ações de combate ao desmatamento. Entretanto, um grande número de cadastros é realizado de maneira errônea gerando dados inconsistentes, levando estes a serem cancelados e/ou a serem pedidas retificações para o devido preenchimento do cadastro. Realizar estas análises, identificando os cadastros preenchidos de maneira incorreta (espúrios) manualmente, possui um grande custo, dada a necessidade de mão de obra especializada, necessitando de um grande período de tempo, devido à imensa quantidade de imóveis rurais no Brasil. Neste contexto, este trabalho tem como objetivo fornecer um sistema inteligente baseado em aprendizagem de máquina que permita verificar e classificar os registros do CAR em registros espúrios e não espúrios (ou cancelados e aprovados) de maneira rápida e eficaz. Para isto, foram aplicadas metodologias que envolvem todo o *pipeline* de uma aplicação envolvendo ciência de dados e aprendizagem de máquina. Desde o pré-processamento, com a limpeza e seleção de atributos, seguido pelo treinamento e validação dos classificadores e, por fim, o uso de algoritmos de aprendizagem de máquina interpretável com o objetivo de avaliar como cada atributo impactou na tomada de decisão pelos classificadores. Foram aplicados 6 modelos de classificação e avaliados seus resultados de acordo com cada formato de pré-processamento, além disto, um modelo de interpretação de classificadores foi utilizado em comparativo com as interpretações internas de modelos que possuem interpretabilidade. Os resultados preditivos mostram índices de desempenho em classificação acima de 90% para todas as medidas de avaliação utilizadas no conjunto de validação e as interpretações elencaram as variáveis que mais influenciam na classificação automática. Assim, o método mostrou-se viável para uma aplicação em um cenário real aplicado ao Cadastro Ambiental Rural.

Palavras-chave: Cadastro Ambiental Rural. Classificação de dados. Dados Desbalanceados. Aprendizagem de Máquina Interpretável. Ciência de dados.

ABSTRACT

The Rural Environmental Registry (CAR) is a mandatory electronic public registry for all rural properties in the Brazilian territory, integrating environmental information from the properties, helping with the environmental monitoring and contributing to actions to combat deforestation. However, a large number of registrations are made erroneously, generating inconsistent data, leading these to be cancelled and/or to request rectifications for the correct completion of the registration. Performing these analyses, identifying the incorrectly completed registries (spurious) manually, has a great cost, given the need for specialized labor, requiring a large amount of time, due to the immense amount of rural properties in Brazil. In this context, this work aims to provide a smart machine learning-based system that allows to check and classify CAR records into spurious and non-spurious (or cancelled and approved) registries in a fast and effective way. To do this, methodologies involving the entire pipeline of an application involving data science and machine learning have been applied. From pre-processing, with attribute cleaning and selection, followed by training and validation of the classifiers, and finally the use of interpretable machine learning algorithms with the goal of evaluating how each attribute impacted the decision making by the classifiers. Six classification models were applied and their results evaluated according to each preprocessing format, and a classifier interpretation model was used to compare the internal interpretations of models that have interpretability. The predictive results show classification performance rates above 90% for all evaluation measures used in the validation set, and the interpretations listed the variables that most influence automatic classification. Thus, the method proved to be viable for application in a real scenario applied to the Rural Environmental Registry.

Keywords: Rural Environmental Registry. Data Classification. Imbalanced Data. Interpretable Machine Learning. Data Science.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

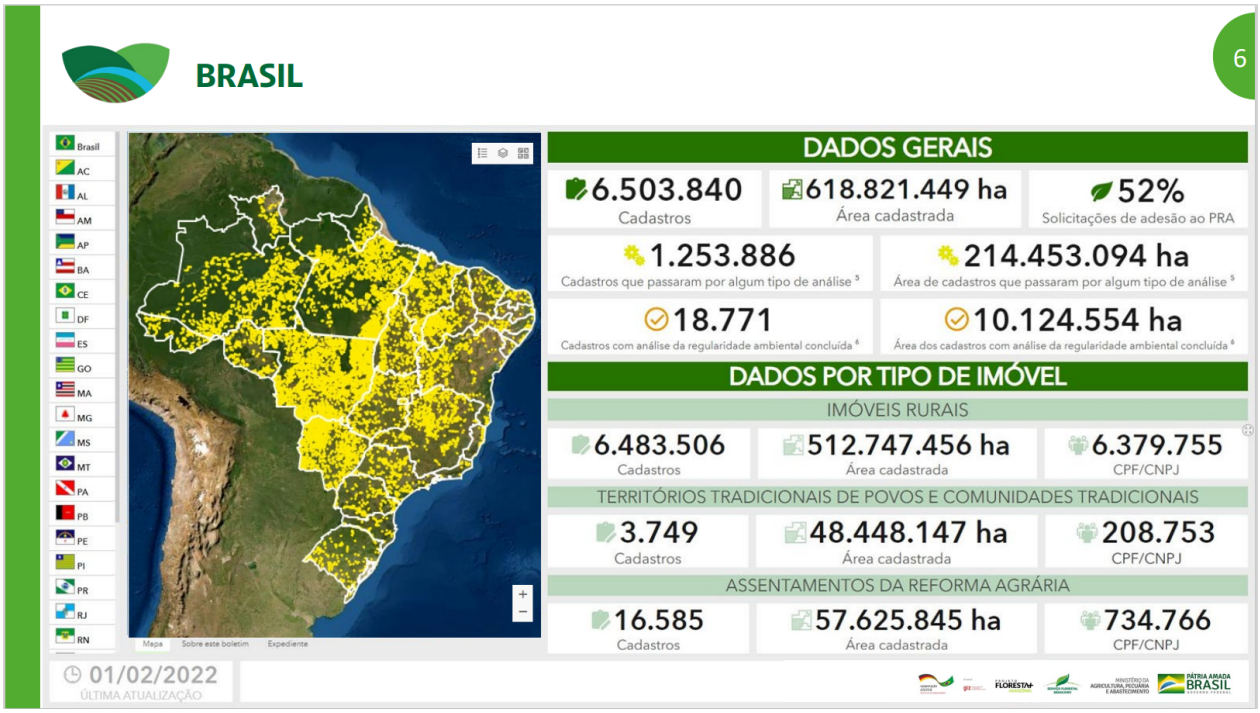
Com o objetivo de monitorar propriedades rurais, auxiliar no combate ao desmatamento e incentivar o devido manejo sustentável das propriedades no campo, o governo brasileiro desenvolveu o Cadastro Ambiental Rural (CAR), inserido no Serviço Florestal Brasileiro (ROITMAN et al., 2018; JUNG et al., 2017; BRASIL, 2012). O cadastro consiste em um registro eletrônico público obrigatório, no qual são reunidas diversas informações dos imóveis rurais no Brasil inseridas pelo cadastrante por meio do SiCAR (Sistema do Cadastro Ambiental Rural). No SiCAR, o cadastrante insere dados geográficos como a localização do terreno, área, feições do terreno (rio, vegetação existente, tipo de vegetação, por exemplo), dentre outros dados. O CAR vem auxiliando no monitoramento e também na investigação de pesquisadores sobre a influência da implementação da plataforma nas ações de desmatamento, invasões de terra, dentre outras irregularidades (L'ROE et al., 2016).

De forma a incentivar os agricultores a fazerem o cadastro e manterem o mesmo devidamente regularizado, o governo brasileiro promove incentivos aos proprietários rurais que realizam o cadastro, como crédito rural facilitado com prazos e taxas melhores que o praticado no mercado, facilitação na contratação de seguro agrícola, dentre outros dispostos no Código Florestal (BRASIL, 2012). Além do cadastro em si, o monitoramento geográfico das áreas ocupadas de forma a mapear as áreas degradadas e auxiliar no combate ao desmatamento é fundamental, conforme recomendam os estudos realizados por Santos et al. (2020) e Arvor et al. (2021).

Além do monitoramento geoespacial das áreas rurais, analisar os registros do CAR de maneira a cancelar um cadastro em caso de eventual irregularidade também contribui para o monitoramento ambiental dos imóveis rurais. Entretanto, tal análise é um grande desafio, dado que o Brasil possui mais de 8 milhões de imóveis rurais, segundo números da Receita Federal (2022), e o CAR, atualmente, conta com mais de 6,5 milhões de cadastros ativos, segundo o último boletim informativo emitido em fevereiro de 2022 pelo Serviço Florestal Brasileiro (2022b). Destes cadastros ativos, menos de 20% passaram por algum tipo de análise até a emissão do boletim, enquanto o número de cadastros com a análise completa é de menos de 1%. Um infográfico mais detalhado sobre os números do CAR pode ser visto na Figura 1.1.

Dados os números e o infográfico supracitados, propor um sistema que possa ser automático ou semiautomático para a realização das análises dos registros, agilizando a tomada de decisão sobre aprovar ou cancelar um cadastro do CAR é de suma importância para a melhoria do sistema do Cadastro Ambiental Rural. Esta importância é válida tanto para o Serviço Florestal Brasileiro otimizar o processo de análise dos

Figura 1.1 – Infográfico sobre os números do CAR lançados no último boletim informativo do Serviço Florestal Brasileiro.



Fonte: Serviço Florestal Brasileiro (2022b).

cadastros, quanto para os agricultores que terão seus cadastros revistos de forma mais rápida, agilizando-os na regularização ambiental. Imerso nesta problemática, este trabalho propõe uma aplicação que possa realizar a classificação automática dos registros do Cadastro Ambiental Rural, utilizando, para isto, algoritmos de aprendizagem de máquina.

1.1 Objetivos

Este trabalho tem como objetivo geral propor uma aplicação de um modelo de aprendizagem de máquina capaz de classificar, com eficiência, os dados do Cadastro Ambiental Rural no que tange presença ou não de informações incorretas, e fornecer uma análise de interpretação da classificação gerada, propondo uma analise detalhada das predições automatizadas. A principal contribuição deste trabalho é a inclusão de um sistema que permita acelerar a tomada de decisão acerca dos registros do CAR, dado que as análises manuais requisitam grande período de tempo para avaliar e classificar todos os cadastros ativos. Como objetivos específicos, tem-se: (i) tratar o conjunto de dados de entrada fornecido para propor um padrão de base de dados; (ii) realizar pré-processamento na base de dados de maneira a selecionar os atributos mais relevantes

à classificação; (iii) implementar classificadores e analisar os resultados gerados por estes; (iv) escolher o modelo de melhor resultado preditivo e realizar testes de interpretação.

1.2 Estrutura do Trabalho

O presente trabalho segue dividido da seguinte forma: no Capítulo 2 será apresentada uma revisão bibliográfica contendo as definições e exemplos de aplicações das técnicas utilizadas neste trabalho; os materiais e métodos aplicados neste trabalho estão inseridos no Capítulo 3; os resultados e as discussões acerca dos mesmos estão situados no Capítulo 4; enquanto no Capítulo 5 estão as conclusões do estudo e os apontamentos para os próximos passos do trabalho.

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo será realizada uma revisão acerca do Cadastro Ambiental Rural, além de uma revisão acerca dos algoritmos aplicados à este trabalho, juntamente com a apresentação de aplicações vistas na literatura. O capítulo está dividido em uma seção abordando o CAR (seção 2.1); seguido pela seção 3.3, onde são apresentados os algoritmos classificadores; a seleção de atributos e o modelo utilizado são apresentados na seção 2.3; na seção 2.4 são apresentadas as técnicas de *oversampling*; finalizando o capítulo, na seção 2.5 é apresentado o assunto de aprendizagem de máquina interpretável e os algoritmos utilizados nesta dissertação.

2.1 Cadastro Ambiental Rural

O Cadastro Ambiental Rural (CAR) é uma plataforma desenvolvida pelo governo brasileiro com o objetivo de monitorar imóveis rurais, permitindo um monitoramento do uso da terra no Brasil e promover práticas de controle de desmatamento e licenciamento ambiental (JUNG et al., 2017) (ROITMAN et al., 2018) (BRASIL, 2012). Além disto, de acordo com o Código Florestal disposto em Brasil (2012), os proprietários que realizam o registro no CAR e o mantém regularizado adquirem benefícios. Dentre estes benefícios, podem ser mencionados: facilidades para aquisição de crédito rural com juros menores, além de prazos e limite maiores; seguro agrícola facilitado; isenção de alguns tributos sobre insumos e equipamentos; preferência na inserção em programas de incentivo ao comércio agrícola, dentre outros dispostos na lei.

Entretanto, mesmo com os benefícios apresentados acima, melhorias no incentivo aos produtores na parte de regularização ambiental para se obter o cadastro ainda se fazem necessárias. A regularização ambiental faz parte dos requisitos para que o cadastro seja devidamente aprovado. Para isto, em caso de pendências dos agricultores neste âmbito, são necessárias compensações de degradação ambiental e/ou recuperação de áreas degradadas. Pacheco et al. (2021) levantaram a discussão sobre o interesse dos agricultores na adesão ao CAR.

Arvor et al. (2021) reportaram sobre outro fator importante para o sucesso do CAR em seu objetivo: o monitoramento geográfico das áreas ocupadas. Uma vez que ocupações anteriores à 2008 foram anistiadas, faz-se importante, atualmente, monitorar novas degradações a fim de localizar os infratores e aplicar as devidas sanções legais. Outro trabalho que reitera o monitoramento espacial aliado ao CAR é apresentado por Santos et al. (2020).

O registro no CAR é feito por um cadastrante, este podendo ser, ou não, o proprietário do imóvel rural. Para a realização do registro, a pessoa precisa baixar o módulo de cadastro no site do SiCAR. Neste site, será

necessário informar a Unidade Federativa (UF, estado) onde a propriedade se encontra e, em seguida, realizar o *download*. Em posse do módulo de cadastro, é necessário baixar as imagens de satélite dos municípios em que estejam inseridos os imóveis rurais. Cabe ressaltar que o módulo permite a realização de vários cadastros por vez, onde, após a finalização dos cadastros desejados, é feito o *upload* de todos juntos para o SiCAR. Este recurso permite a realização do cadastro de maneira *offline*, uma vez que há áreas rurais com baixa cobertura de *internet*. Realizados os procedimentos, o cadastrante pode realizar os cadastros, onde são inseridas informações sobre o cadastrante, o imóvel, o proprietário, a documentação, o mapa geográfico da propriedade e o questionário acerca desta. A tela inicial do módulo de cadastro do CAR pode ser vista na Figura 2.1.

Figura 2.1 – Tela inicial do módulo do Cadastro Ambiental Rural

Fonte: Serviço Florestal Brasileiro (2021).

2.2 Algoritmos de classificação

A classificação de dados ou classificação de padrões é uma tarefa de mineração de dados que consiste em predizer um determinado rótulo para um conjunto de dados (HAN; PEI, 2011; EVSUKOFF, 2020). Trata-se de uma tarefa similar à regressão, com a diferença na saída: enquanto a saída de um modelo regressor consiste em um valor numérico contínuo, a saída de um classificador consiste em uma saída categórica. Não obstante, grande parte dos algoritmos de aprendizagem que são aplicados em problemas de classificação

podem ser adaptados para problemas de regressão e vice-versa. Nas subseções a seguir serão apresentados os modelos de classificação utilizados neste trabalho.

2.2.1 Regressão Logística

A Regressão Logística ou modelo *logit* consiste em um modelo de classificação baseado na função logística para estimação da probabilidade *a posteriori* de uma determinada observação pertencer à determinada classe. Para isto, são utilizadas funções lineares, tomando o vetor de entrada x como variável independente (HASTIE et al., 2009b; EVSUKOFF, 2020). O modelo logístico consiste na aplicação da função sigmoide (logística ou *logit*) ao modelo linear, cuja função de separação, em problemas de classificação binária, é definida pela equação 2.1 (EVSUKOFF, 2020):

$$g(\mathbf{x}, \theta) = \frac{1}{1 + exp(-\hat{\mathbf{x}}\theta^T)} \tag{2.1}$$

onde $\hat{\mathbf{x}} = [1, \mathbf{x}]$ é o vetor de regressores, similar a uma regressão linear e θ é o vetor de parâmetros do modelo.

O ajuste dos parâmetros θ se dá pela máxima verossimilhança, onde se tem por objetivo maximizar o potencial do modelo de representar corretamente os dados que o projeta. De maneira a simplificar o processamento, é utilizado o negativo do logaritmo da função de verossimilhança, de maneira que se possa transformar em um problema de otimização para o ajuste de parâmetros do modelo. Para problemas de duas classes, o ajuste é dado pela equação 2.2:

$$l(\theta) = -\sum_{i=1}^{N} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) \tag{2.2}$$

onde N é o número de observações da base de dados de treinamento do modelo, y_i é a saída real da i-ésima observação e $\hat{y}_i$ é a saída estimada pelo classificador da i-ésima observação.

Assim como outros modelos lineares, a Regressão Logística tem um baixo custo computacional, dada a simplicidade de sua formulação. Juntamente com sua simplicidade, outra característica importante que o modelo logístico possui é sua capacidade de ser interpretável. Tal propriedade pode ser observada por meio dos pesos do modelo ajustados durante o treinamento, os quais refletem a influência de cada atributo na saída do classificador. O sinal do peso indica para qual classe o atributo está influenciando e o valor do peso indica o quão grande é esta influência.

Um exemplo visto na literatura é o trabalho reportado por Robles-Velasco et al. (2020), onde os autores aplicaram Regressão Logística para detecção de falhas na tubulação de uma rede de abastecimento de

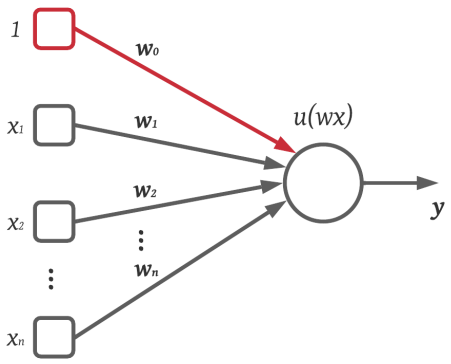
água. Os resultados preditivos mostraram uma acurácia da Regressão Logística ligeiramente superior à de um modelo de SVM (76,9% do modelo logístico contra 75,0% do SVM). Além do resultado preditivo, os autores utilizaram as propriedades da Regressão Logística para identificar como cada variável impacta na incidência de falhas como, por exemplo, foi identificado neste estudo de caso que os tubos de menor diâmetro eram mais propensos à falha.

2.2.2 Redes Neurais Artificiais

As Redes Neurais Artificiais são modelos matemáticos baseados no princípio do funcionamento dos neurônios biológicos (HAYKIN, 2007) e largamente utilizados em suas variações em diversos problemas de aprendizagem de máquina.

O primeiro modelo de uma Rede Neural foi desenvolvido por McCulloch e Pitts (1943) onde os autores propuseram um modelo matemático baseado no neurônio biológico, onde cada entrada possuía um respectivo peso e a saída se dava por uma função degrau. Um grafo representando um neurônio artificial pode ser visto na Figura 2.2.

Figura 2.2 – Exemplo em grafo do Neurônio Artificial proposto por McCulloch e Pitts (1943), $u(wx)$ representa a função degrau da entrada x multiplicada pelo peso w.



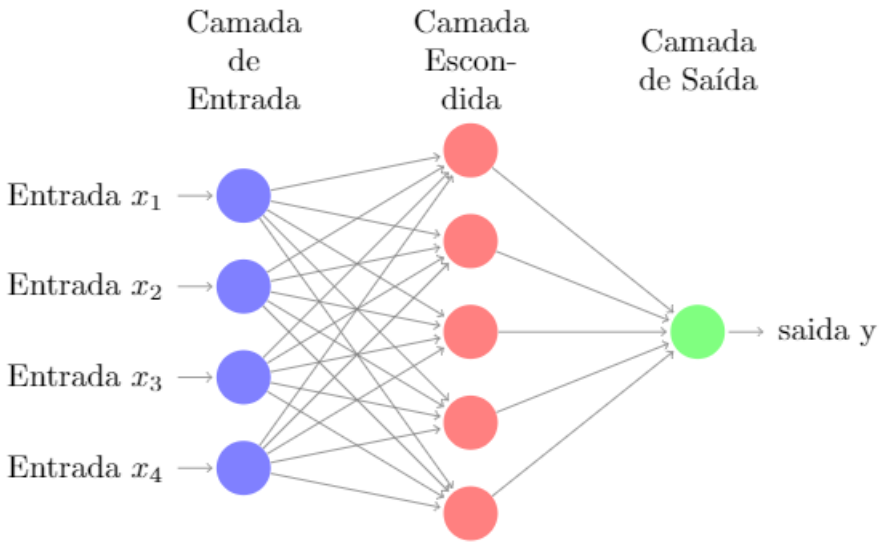
Fonte: Do Autor (2021).

Após o desenvolvimento do primeiro modelo do Neurônio Artificial, dois trabalhos tiveram grande importância para o aprimoramento das Redes Neurais. Estes trabalhos foram o aprendizado Hebbiano (HEBB, 1949) e o *perceptron* de Rosenblatt (ROSENBLATT, 1958), onde foi proposto o *perceptron* como um modelo de aprendizado de máquina, capaz de simular funções lógicas como AND e OR. Tais desenvolvimentos

precedem o conceito de Rede Neural Artificial e são o início das Redes Neurais conhecidas e utilizadas atualmente.

Seguindo os postulados propostos pelos autores supracitados e, com o aperfeiçoamento dos métodos de ajuste das Redes Neurais, novas topologias foram desenvolvidas, dentre estas a Rede Neural de *perceptron* multicamadas (MLP, do inglês, *Multi-Layer Perceptron*) (HAYKIN, 2007). As redes MLP são do tipo *feedforward*, onde o fluxo de sinal ocorre unidirecionalmente da camada de entrada para a camada de saída. Outra característica das MLP está associada às conexões entre os neurônios da rede, podendo a rede ser totalmente ou parcialmente conectada. Uma rede totalmente conectada é aquela onde os pesos sinápticos possuem valores não nulos, enquanto uma rede MLP parcialmente conectada possui pesos anulados durante sua implementação. Um exemplo de um grafo de Rede *feedforward* totalmente conectada pode ser visualizado na Figura 2.3.

Figura 2.3 – Grafo de uma Rede Neural Muticamadas do tipo *feedforward* totalmente conectada.



Fonte: Do Autor (2021).

Tomando o grafo da Figura 2.3 como exemplo, o funcionamento de uma Rede Neural dá-se pela soma dos pesos sinápticos ajustados pelo treinamento tendo a adição de uma constante de *bias* também ajustada. Após esta operação linear, o resultado gerado passa por uma função de ativação, gerando a não-linearidade da Rede Neural. Este processo é realizado em cada camada até a saída, gerando o valor categórico de classe ou a probabilidade de pertencimento de uma amostra a cada classe tida na saída (em problemas de classificação) ou o valor da saída da função na qual a rede foi utilizada para aproximar (em problemas de regressão).

O número de neurônios tanto na camada de entrada, quanto na camada de saída são fixos, o primeiro é dado pela dimensionalidade da matriz de entrada (número de atributos do conjunto de dados) enquanto o segundo é determinado ou pelo número de classes para classificação ou é unitário, para os problemas de regressão, onde só há uma saída por amostra. O número de neurônios das camadas intermediárias e o número de camadas intermediárias é estipulado pelo usuário. A operação realizada em cada camada da Rede Neural pode ser escrita de maneira generalizada na forma matricial, relacionando a saída com a entrada de uma determinada camada k da Rede Neural. Esta função é definida pela equação (2.3):

$$\mathbf{u}_k = f(\mathbf{W}_k \mathbf{u}_{k-1} + \mathbf{b}_k) \tag{2.3}$$

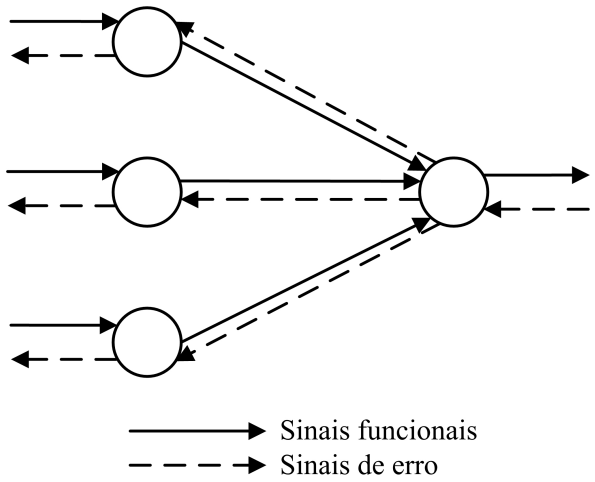
sendo u a saída de uma camada k, $\mathbf{W}_k$ é a matriz de pesos da camada k; $\mathbf{b}_k$ são os *bias* da camada e $\mathbf{u}_{k-1}$ à saída da camada anterior $(k-1)$, sendo que, para $k-1=0$, $\mathbf{u}_{k-1} = \mathbf{x}$, sendo $\mathbf{x}$ o vetor contendo os dados de entrada. A função $f(.)$ refere-se à função de ativação da Rede Neural.

Existem vários tipos de funções de ativação para Redes Neurais, podendo estas serem lineares, ou não. As funções de ativação mais comuns utilizadas em Redes Neurais são: função logística (ou sigmoide), função linear retificada (ReLu) e tangente hiperbólica.

O processo de treinamento das redes MLP, ou seja, o ajuste dos pesos sinápticos é feito por meio do algoritmo de retropropagação do erro (ou *error backpropagation*, do inglês, ou, somente, *backpropagation*) (HAYKIN, 2007). O *backpropagation* tem seu funcionamento da seguinte forma: enquanto o sinal funcional da rede percorre da camada de entrada em direção à camada de saída, o sinal de erro de estimação da rede percorre no sentido contrário ao sinal funcional. Desta forma, partindo deste fluxo do sinal de erro, é possível ajustar os pesos da Rede Neural por meio do método do gradiente descendente. Desta forma, é possível ajustar os pesos de todos os neurônios em todas as camadas da rede. Na Figura 2.4 é apresentado um exemplo do funcionamento do *backpropagation*. Baseando-se no *backpropagation*, outros algoritmos de otimização para os pesos da rede foram desenvolvidos, além do gradiente descendente, como, por exemplo: o gradiente estocástico descendente (sgd) (AGGARWAL, 2018), o gradiente estocástico adaptativo (adam) (KINGMA; BA, 2015) e o algoritmo de ajuste do tipo quasi-Newton (lbfgs) (LIU; NOCEDAL, 1989).

Aplicações envolvendo Redes Neurais MLP podem ser vastamente encontrados na literatura, como no estudo reportado por Karimi e Heidarian (2021), onde foi utilizada uma Rede MLP adaptada (parcialmente conectada) para detecção de emoções por meio de imagens de rosto. Os autores utilizaram polinômios de Legendre para extrair características de regiões da face como a boca, os olhos e as sobrancelhas. Os

Figura 2.4 – Grafo ilustrando o funcionamento do *backpropagation* e seus fluxos de sinais



Fonte: Do Autor (2021).

pesquisadores reportaram uma acurácia média entre as classes de 96% e resultados de *recall* variando entre 75% até 100%.

Outro exemplo, pode ser vista no trabalho desenvolvido por Pizzaia et al. (2018), no qual os autores utilizaram uma Rede Neural MLP para classificação da qualidade dos grãos de café por meio de reconhecimento de padrões em imagens. Foram utilizados como atributos, características dos grãos como cor (taxas dos valores RGB), taxa de arredondamento e área do grão para a classificação. O modelo proposto atingiu uma acurácia média de 94% para a tarefa, conseguindo classificar os grãos em alta e baixa qualidade.

2.2.3 *Random Forests*

Florestas Aleatórias (ou *Random Forests*, do inglês) são um modelo de aprendizagem por *ensemble* ou comitê. Ou seja, constituem-se por um conjunto de modelos de aprendizagem de maneira que possuam um maior poder de discriminação (BREIMAN, 2001). As *Random Forests* fazem uso de modelos de Árvores de Decisão, algoritmo clássico de aprendizagem de máquina, de arquitetura simples e treinamento rápido.

O modelo treina um determinado número de Árvores de Decisão, no qual os hiperparâmtros das árvores como profundidade máxima, o índice de qualidade do particionamento da árvore (entropia ou índice Gini), tamanho do particionamento, etc., são especificados pelo usuário. Assim que treinadas as árvores, a classificação da *Random Forest* se dá pela votação, onde a classe com o maior número de votos é atribuída à amostra de entrada. O treinamento de cada árvore se dá por um subconjunto do conjunto total de dados, onde cada subconjunto é selecionado de maneira aleatória, tanto em amostras, quanto em atributos. Cade salientar

que o processo de amostragem dos subconjuntos é realizado com repetição (*bootstrap*). A função de margem dada pelo *Random Forest* pode ser dada pela equação (2.4):

$$mg(\mathbf{X},Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y}(avI(h_k(\mathbf{X}) = j)), \tag{2.4}$$

onde h_k se refere ao k-ésimo classificador por Árvore de decisão, $\mathbf{X}$ é o conjunto de dados de entrada, Y é a saída e $I(.)$ é a função indicadora.

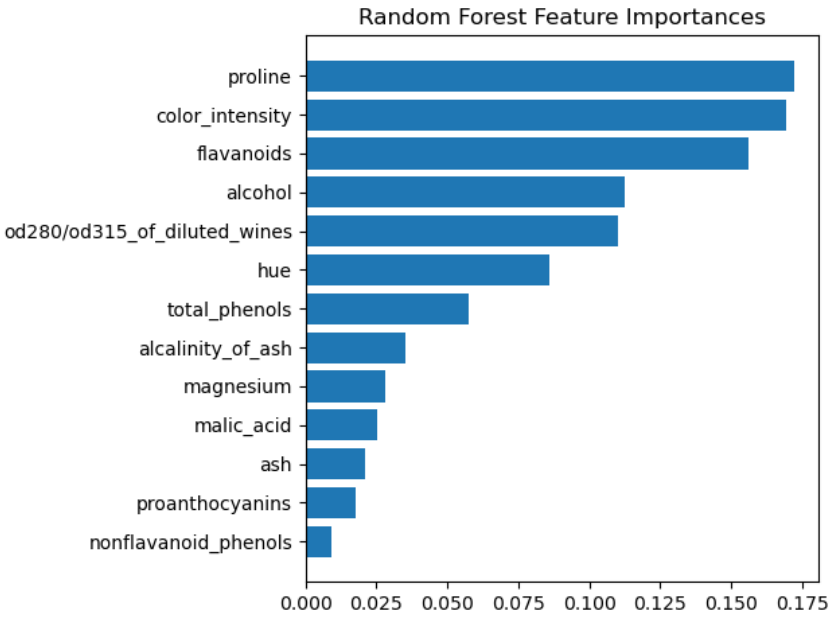
As *Random Forests* possuem aplicações em diversas áreas com bons resultados. Kalaiselvi e Thangamani (2020) realizaram um estudo utilizando *Random Forests* como ferramenta para auxiliar na identificação de padrões em proteínas, onde o modelo reduziu os erros de classificação. Os autores utilizaram uma base de dados referente à uma sequência de aminoácidos com atributos extraídos utilizando a correlação de Pearson ponderada. Os resultados de classificação em duas bases de dados mostraram um desempenho superior da *Random Forest* em relação aos demais classificadores: para a base de dados *VariBench*, a *Random Forest* atingiu uma acurácia de 90% contra 84% do segundo melhor resultado no comparativo, para a base de dados *PDB*, a acurácia da *Random Forest* foi de 96% contra 87% do segundo melhor classificador.

Em outro estudo, foi visto que as *Random Forests* podem gerar resultados compatíveis com modelos mais sofisticados, como as Redes Neurais profundas (*Deep Learning*) em problemas de classificação, como o estudo realizado por Sagayaraj, Jithesh e Roshani (2021). Neste trabalho, os autores utilizaram sinais sintéticos de assinatura de radar para diversos ângulos de azimute com adição de ruído gaussiano ao sinal. Os resultados apresentados mostram resultados próximos entre os modelos profundos utilizados como Redes Convolucionais (CNN) e Redes LSTM e a *Random Forest*. Para alguns valores de SNR, a *Random Forest* superou os modelos profundos, como no SNR igual a 9dB, onde os resultados para a *Random Forest*, Rede CNN e Rede LSTM foram, respectivamente, 97,69%, 94,47%, 95,39%, onde a diferença chegou à 3,22% favorável para a *Random Forest*. A maior diferença contra a *Random Fores* foi de 5,53%, para os demais valores de SNR, a diferença entre as acurácias permaneceu entre os dois extremos.

Outro ponto interessante sobre as *Random Forests* é a sua capacidade de interpretação, onde é possível analisar a relevância que cada atributo impacta sobre a classificação. Tal propriedade é importante para analisar quais variáveis possuem maior impacto na classificação do ponto de vista do classificador e confrontar com as análises vistas pela aplicação. Desta forma, a interpretação gerada pelo modelo de aprendizagem pode contribuir com as análises práticas, comprovando interações entre os atributos de entrada e saída já conhecidos e sugerir novas interações entre estes. Um exemplo gráfico de visualização da importância das

variáveis utilizando *Random Forest* é apresentada na Figura 2.5, onde é aplicada uma *Random Forest* na base de dados *wine* (LICHMAN et al., 2022).

Figura 2.5 – Exemplo do uso de *Random Forests* na obtenção da importância dos atributos para classificação utilizando a base de dados *wine*.



Fonte: Do Autor (2022).

Aplicações envolvendo *Random Forest* para classificação em conjunto com a interpretação podem ser encontradas na literatura, como no trabalho reportado por Zhong, Song e Yang (2019), no qual os autores aplicaram *Random Forest* para a classificação de navios utilizando imagens de satélite como base de dados. Como resultados, os autores obtiveram uma acurácia de 86,5% em dados de teste. Além do resultado preditivo, foi observado que as variáveis de maior importância foram o comprimento do navio e o área da embarcação. Também foi testada uma *Random Forest* utilizando somente as duas variáveis mais importantes, gerando uma acurácia em teste de 73,3%. Estes exemplos mostram que a *Random Forest* pode, conjuntamente fornecer boa capacidade preditiva e interpretabilidade.

2.2.4 Algoritmos de *Boosting*

Baseando-se na ideia de aprendizagem por *ensemble* também presente nas *Random Forests*, outros algoritmos fazem uso de múltiplos modelos de aprendizagem de menor complexidade, dentre estes, o

AdaBoosting e o *Gradient Boosting*. Diferentemente do primeiro citado, os algoritmos apresentados nesta seção podem fazer uso de outros modelos além das Árvores de Decisão.

Além da possibilidade do uso de diferentes algoritmos de aprendizagem, outra diferença entre o algoritmo descrito na seção anterior e os dois apresentados nesta seção se dá pelo formato do comitê e a geração da classificação pelo modelo de *ensemble*. Enquanto o primeiro faz uso de amostragem e do próprio *ensemble* por meio do *bagging*, onde todos os classificadores possuem o mesmo peso na classificação final, os dois últimos fazem uso do *boosting*, onde tanto a reamostragem, quanto a classificação dos dados se dá com pesos variáveis tanto para as amostras quanto para os modelos individuais dentro do comitê.

A regra de classificação do *ensemble* utilizando *boosting* se dá pela média ponderada da classificação gerada pelas Árvores, onde cada uma possui um determinado peso na classificação de acordo com o erro gerado durante o treinamento, onde amostras com classificação errada têm maior peso em comparação com as amostras cujo o classificador não tem errado a predição. Em outras palavras, o *boosting* dá maior importância aos eventos de maior dificuldade em serem preditos corretamente em detrimento da menor importância para as amostras em que o modelo não possui erros de predição.

A partir do conceito de *boosting*, métodos de treinamento de *ensembles* foram desenvolvidos, cada um com sua diferença no processamento. A principal diferença encontra-se no ajuste dos pesos para o comitê. O *AdaBoost* utiliza uma função pré-determinada para o ajuste de pesos, a depender de cada algoritmo (HASTIE et al., 2009a), enquanto o *Gradient Boosting* ajusta os pesos pelo método do gradiente descendente (FRIEDMAN, 2002).

O ajuste dos pesos do *ensemble* do modelo *AdaBoost* é dado pelo algoritmo SAMME.R, cuja formulação é definida pela equação (2.5).

$$w_i \leftarrow w_i \cdot exp\left(\frac{K-1}{K}\mathbf{y}_i^{\intercal} log(\mathbf{p}^{(m)}(\mathbf{x}_i))\right), i = 1,...,n, \tag{2.5}$$

onde w_i é o i-ésimo peso amostral para a respectiva i-ésima amostra, K o número de classes da base de dados, $\mathbf{y}_i$ é o valor da i-ésima saída do banco de dados de treinamento e $\mathbf{p}(\mathbf{x}_i)$ refere-se à probabilidade da i-ésima entrada do banco de dados x pertencer à determinada classe.

O *Gradient Boosting* ajusta seus pesos por meio de uma aproximação do gradiente descendente, tal aproximação procura minimizar uma dada função custo Ψ. Portanto, a função de ajuste dos pesos do modelo, denominado por γ_{lm} é descrita pela equação (2.6).

$$\gamma_{lm} = \arg\min_{\gamma} \left(\sum_{\mathbf{x}_{\pi(i)} \in R_{lm}} \Psi(y_{\pi(i)}, F_{m-1}(\mathbf{x}_{\pi(i)}) + \gamma) \right), \tag{2.6}$$

onde $F(x)$ representa a função de decisão do modelo que realiza o mapeamento da entrada $\mathbf{x}$ com a saída y, R_{lm} é a região contendo uma subamostragem do conjunto de dados de treinamento e $\pi(i)$ é o i-ésimo valor da permutação dentro do conjunto de treinamento.

Na literatura, é possível encontrar exemplos envolvendo os dois algoritmos de *boosting* apresentados nesta seção, como o estudo realizado por Uddin, Fatema e Dhar (2020), onde os autores utilizaram o *AdaBoost* para predição de riscos de depressão em colaboradores da área de tecnologia. Os autores, juntamente com um psiquiatra, elaboraram um questionário a ser preenchido pelos colaboradores. Em seguida, os dados passaram por um procedimento de limpeza e aplicados a modelos de classificação. Como resultados, o *Adaboost* utilizando árvores de decisão como classificador de base, atingiu os melhores desempenhos preditivos, com 97,4% de acurácia, tendo o segundo melhor classificador obtido 96,5%.

Em Brandão et al. (2019) os autores aplicaram o *AdaBoost* na classificação de desempenho escolar (aprovação ou reprovação), utilizando como base de dados, perfis de uso de uma plataforma de ensino virtual. Foi realizado um procedimento de extração de características antes do uso do classificador e, como resultados preditivos, foi obtido um índice *F1-Score* de 90%, superando os estudos anteriormente realizados sobre a mesma aplicação utilizando a mesma base de dados, que tinham obtido um *F1-Score* de 83%.

Aplicações envolvendo o *Gradient Boosting* podem ser vistas em diferentes áreas de estudo, como em Alsirhani, Sampalli e Bodorik (2018), onde foi aplicado o modelo de *Gradient Boosting* para detecção de falhas em redes do tipo negação de serviço distribuída. Como base de dados, foram utilizados dados reais de tráfego de rede, consistindo em vetores binários, correspondendo aos pacotes de dados que transitam pela rede ao longo do tempo. Os resultados de acurácia atingiram, em alguns casos, 97,1% de acurácia.

Outra aplicação pode ser vista no trabalho realizado por Li et al. (2021), no qual os autores aplicaram o *Gradient Boosting* na predição de mortalidade em pacientes internados por sépsis. Foi utilizada uma base de dados pública contendo diversas variáveis sobre os pacientes, como idade, tempo de internação, sinais vitais como saturação de oxigênio, dentre outras. Os reultados de classificação, envolvendo comparativo com outros classificadores, apontaram o *Gradient Boosting* como o de melhor desempenho, com 95,4% de acurácia, tendo o segundo melhor classificador atingido 93,8%. Além do resultado de classificação, os autores utilizaram a propriedade do *Gradient Boosting* de calcular a importância dos atributos para identificar os atributos de maior relevância para a classificação.

2.2.5 Máquinas de Vetor de Suporte

As Máquinas de Vetor de Suporte (SVM, do inglês *Support Vector Machines*) são modelos de aprendizagem de máquina baseados na teoria do aprendizado estatístico, fundamentada por Vapnik (1998). O SVM busca um hiperplano ótimo em que haja a maximização da distância entre as classes. Quando os problemas são linearmente separáveis, o algoritmo separa por meio da obtenção deste hiperplano.

Em conjuntos de dados não linearmente separáveis, o SVM realiza uma transformação no espaço das variáveis, elevando a dimensionalidade do problema por meio das funções de *kernel*. Desta forma, na dimensão transformada pelo *kernel*, é possível obter fronteiras de separação lineares. As funções geralmente utilizadas como *kernel* para o SVM são: linear, polinomial, função de base radial (RBF, do inglês *Radial Basis Function*) e tangente hiperbólica (HASTIE et al., 2009b). As funções de *kernel* estão descritas, respectivamente, nas equações 2.7, 2.8, 2.9 e 2.10.

$$k(\mathbf{x},\mathbf{x}') = \langle \mathbf{x},\mathbf{x}' \rangle \tag{2.7}$$

$$k(\mathbf{x},\mathbf{x}') = (\gamma \langle \mathbf{x},\mathbf{x}' \rangle + r)^d \tag{2.8}$$

$$k(\mathbf{x},\mathbf{x}') = exp(-\gamma \left\| \mathbf{x}-\mathbf{x}' \right\|)^2 \tag{2.9}$$

$$k(\mathbf{x},\mathbf{x}') = tanh(\gamma \langle \mathbf{x},\mathbf{x}' \rangle + r) \tag{2.10}$$

onde $\mathbf{x}$ e $\mathbf{x}'$ são o vetor de dados de entrada mapeados no novo espaço de parâmetros, γ é um hiperparâmetro especificado pelo usuário (por padrão, geralmente, igual a $\frac{1}{2\sigma^2}$, onde σ^2 é a variância do conjunto de dados de entrada), d é o grau do polinômio do *kernel* polinomial e r é um termo independente. As operações $\langle \mathbf{x},\mathbf{x}' \rangle$ e $\|\mathbf{x}-\mathbf{x}'\|$ se referem, respectivamente, ao produto interno e ao módulo de $\mathbf{x}$ e $\mathbf{x}'$.

O SVM, pela sua construção inicial, é livre de hiperparâmetros, ou seja, não depende de parâmetros estipulados pelo usuário, apenas do próprio ajuste dos parâmetros internos, entretanto, tal situação se restringe somente à conjuntos de dados separáveis, linearmente ou não (EVSUKOFF, 2020). Para conjuntos de dados não separáveis, ou seja, onde não é possível a confecção de uma fronteira de decisão que separe totalmente as classes, que consistem no tipo mais comum de dados em cenário real, o SMV inicial não pode ser aplicado.

Para contornar este problema, foi desenvolvida uma versão do SVM que permite a inclusão de variáveis de folga, que flexibilizam a margem de separação. Assim, a equação de ajuste do modelo SVM, com a inclusão das variáveis de folga, é descrita pela equação 2.11:

$$\begin{gathered} \min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{\lambda}{N}\sum_{i=1}^{N}\xi(i)^k \\ sujeito\ a: \xi(i) =\geq 0;\ v(i)g(\mathbf{x}(i)) \geq 1-\xi(i) \end{gathered} \tag{2.11}$$

onde ξ são as variáveis de folga, N o número de instâncias, λ o parâmetro de regularização e $\mathbf{w}$ a margem de separação do classificador.

Uma aplicação do uso do SVM pode ser vista no estudo reportado por Jose et al. (2021), no qual os autores investigaram o uso sinais de um sistema hidráulico como atributos e o SVM para investigação de vazamento precoce no sistema com uma acurácia geral de 97.5%.

2.3 *Algoritmos de seleção de atributos*

Muitos conjuntos de dados reais possuem centenas ou até, em alguns casos, milhares de atributos em sua base. Treinar classificadores utilizando todos estes atributos, além de demandar grandes recursos computacionais, alta capacidade de processamento e memória, além de um elevado tempo de processamento, não necessariamente geram resultados ótimos. Baseando-se neste pressuposto, selecionar corretamente os atributos da base de dados para a classificação pode, conjuntamente, reduzir a demanda computacional e melhorar os resultados preditivos para a classificação.

Selecionar quais variáveis geram os melhores resultados manualmente é uma tarefa inviável, uma vez que o número de combinações será numeroso. Tendo em vista tal problema, modelos de seleção de atributos foram desenvolvidos com a finalidade de auxiliar nesta escolha e automatizar a sumarização das variáveis de maior importância para a classificação. Os métodos de seleção de atributos podem ser divididos em duas categorias: filtro e *wrapper* (AGGARWAL et al., 2015). Os seletores de atributos do tipo filtro são compostos por algoritmos que computam pesos que demonstram a influência de cada atributo de entrada na classificação. Tais cálculos pode ser por medidas estatísticas (média, variância, etc.), medidas de entropia ou informação mútua, por exemplo. Os modelos de seleção de atributos do tipo *wrapper* utilizam um classificador no processo de seleção e, geralmente, buscam otimizar algum índice de desempenho do classificador como maximizar a acurácia ou minimizar o erro de predição. A diferença principal entre os dois tipos de seletores se dá pelo uso de um classificador de referência e ao subconjunto obtido. Os modelos do tipo filtro fazem

uso somente das equações de obtenção dos pesos e geram um ranqueamento dos atributos, cabendo ao usuário definir qual o melhor subconjunto (quantos atributos utilizar). Os modelos *wrapper* utilizam um classificador de base para a otimização desejada, gerando, na saída, o subconjunto ótimo contendo os atributos selecionados. Logo, os seletores de atributos do tipo filtro possuem menor custo computacional, enquanto os do tipo *wrapper* tendem a gerar os melhores subconjuntos de dados, juntamente com o resultado preditivo.

2.3.1 Discriminante Linear de Fisher

O Discriminante Linear de Fisher ou a Razão de Discriminante de Fisher ou (em inglês, *Fisher's Discriminant Ratio* - FDR) (DUDA; HART, 2001), é um algoritmo linear que pode ser utilizado, tanto para classificação de padrões, quanto para seleção de atributos. O FDR tem como principal característica a minimização das distância intra-classe juntamente com a maximização da distância inter-classe. Em suma, o FDR busca compactar uma determinada classe em uma determinada região, enquanto aumenta a distância entre classes diferentes. Tal raciocínio pode ser aplicado em ambas as tarefas, seja para classificação ou para seleção de atributos. Para o segundo caso, o FDR aplica o conceito de seleção de atributos por meio de um fator J que expressa tal característica, em que, quanto maior o valor do índice J, maior a relevância de tal atributo para a classificação. A formulação do FDR para seleção de atributos em problemas de duas classes é descrita pela equação 2.12 Duda e Hart (2001):

$$\mathbf{J} = (\mu_i - \mu_j)^2 \odot \left(\frac{1}{\sigma_i^2 + \sigma_j^2} \right), \tag{2.12}$$

onde J é o vetor que mostra a relevância de cada atributo para a classificação, μ_i e μ_j são os vetores de médias para as respectivas classes i e j, σ_i^2 e σ_j^2 são as variâncias, respectivamente, das classes i e j e $\odot$ refere-se ao produto ponto a ponto.

O FDR possui como vantagem o baixo custo computacional para calcular o índice J, dado que é somente a aplicação direta da fórmula descrita na equação 2.12. Além disto, trabalhos que utilizaram o FDR obtiveram boa seleção de atributos e resultados de classificação com altas taxas de acerto, como em Barbosa et al. (2016), onde o uso do FDR para selecionar atributos para detecção e classificação de falhas em uma viga engastada. O uso do seletor de atributos reduziu a dimensionalidade do conjuntos de dados, em situações variando o número de atributos de 250 até 1875, para somente 2 atributos com resultados de acerto variando entre 88% até 100%.

2.4 Algoritmos de *Oversampling*

As base de dados reais, geralmente possuem desbalanceamento de classes, por exemplo, bases de dados envolvendo fraudes em transações financeiras tendem a possuir um número significativamente maior de transações normais em comparação ao número de transações fraudulentas. Esta diferença, durante o treinamento de um modelo supervisionado, como um classificador, pode gerar um modelo tendencioso à classe mais numerosa, dado o princípio de ajuste de parâmetros ser baseado, geralmente, no erro médio quadrático, não considerando o erro por classe. Logo, o classificador pode ter uma elevada acurácia média com baixa taxa de acerto da classe menos numerosa.

Com o objetivo de sanar o problema descrito acima, existem abordagens que auxiliam a mitigar o erro da classe minoritária, dentre estes, a ponderação da função de custo (*loss*) ou a geração de dados sintéticos para a classe minoritária. Dente os algoritmos de *oversampling*, podem ser mencionados a superamostragem aleatória (MENARDI; TORELLI, 2014), que consiste na reamostragem aleatória com reposição da classe minoritária *booststrapping* e o algoritmo SMOTE.

2.4.1 SMOTE - *Synthetic Minority Over-sampling Technique*

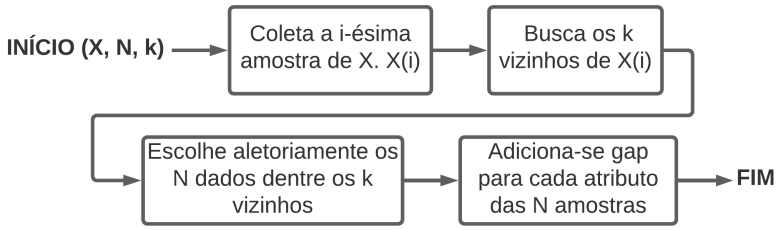
O SMOTE consiste em uma técnica de geração de dados sintéticos com base em um determinado conjunto de dados real, com a finalidade de obter um maior balanceamento de classes em problemas de classificação (CHAWLA et al., 2002). O algoritmo possui como base o k-vizinhos mais próximos, onde ele captura os vizinhos mais próximos de cada amostra e, para os vizinhos escolhidos como amostra, os valores destes são adicionados de um *gap* (valor aleatório entre 0 e 1). Assim, os novos dados sintéticos são gerados.

Um diagrama em blocos do método pode ser visto na Figura 2.6, em que $\mathbf{X}$ é o conjunto de dados reais, N é o valor inteiro de aumento em dados sintéticos e k é o número de vizinhos mais próximos para busca. Assim, o algoritmo obtém novas amostras a partir do conjunto de dados real, mantendo sua distribuição e equilibrando o número de amostras por classe.

Um exemplo gráfico pode ser mostrado na Figura 2.7, onde é apresentado um conjunto de dados aleatório com e sem o uso do SMOTE para *oversampling*. No comparativo, pode ser observada a manutenção da distribuição do perfil dos dados da classe superamostrada.

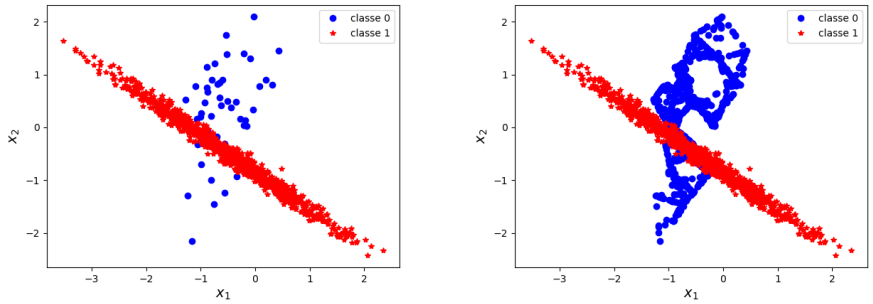
Um maior equilíbrio entre as classes tende a equilibrar os resultados em problemas de classificação, onde o erro acumulado de ambas as classes passam a ter o mesmo peso durante o processo de treinamento.

Figura 2.6 – Diagrama em blocos do funcionamento do SMOTE.



Fonte: Do Autor (2021).

Figura 2.7 – Exemplo do uso do SMOTE em um conjunto de dados de duas dimensões.



(a) Conjunto de dados original (b) Conjunto de dados superamostrado por SMOTE

Fonte: Do Autor (2021).

Exemplos na literatura que fazem o uso do SMOTE mostram que a técnica auxilia na revocação das classes pelo classificador em classes minoritárias cujo número de amostras foi aumentado usando o algoritmo.

Gicić e Subasi (2019) usaram o SMOTE para sobreamostrar a classe minoritária de um conjunto de dados tabular de microcrédito, oriundas de micro e pequenas empresas. A aplicação do SMOTE gerou um aumento na acurácia da classe minoritária de 0,948 para 0,961 e o aumento na acurácia geral do classificador de um valor máximo de 0,946 para 0,966.

Jonathan, Putra e Ruldeviyani (2020), em um trabalho de classificação de dados textuais de uma mídia social, no qual buscava identificar postagens de venda, prática proibida pelos termos da rede social estuda. A classe de postagens contendo venda de produtos é significativamente menor que as demais postagens, logo, sua recuperação se torna mais difícil para os classificadores. O uso do SMOTE permitiu um aumento no desempenho dos classificadores de Regressão Logística e SVM. Na medida de precisão, o aumento foi de 2.59% (de 90.89% para 93.48%) para o modelo logístico e de 5,14% (de 89,06% para 94,20%) para o SVM.

Na medida de revocação (*recall*), o modelo logístico teve um aumento de 9,48% (de 84,85% para 94,33%) e o classificador SVM teve um incremento de 5,05% (de 86,62% para 91,66%).

Os resultados reportados mostraram a utilidade do SMOTE para equilibrar as amostras por classe. Tais casos mostram que fazer o *oversampling* pode auxiliar na melhoria do desempenho de classificadores em problemas envolvendo desbalanceamento.

2.5 Aprendizagem de máquina interpretável

O processo de mineração de dados e aprendizagem de máquina vêm passando por diversas atualizações e apontamentos de melhorias, as quais surgem com o objetivo de aperfeiçoar o procedimento (WANG; CAO; YU, 2020), (GRISCI; KRAUSE; DORN, 2021). Dentre os apontamentos apontamentos sugeridos, o uso de modelos de aprendizagem que podem ser interpretáveis ou algoritmos de interpretação de modelos tradicionais "caixa preta", vêm sendo levantados atualmente. Tal interpretação pode melhor auxiliar o usuário na tomada de decisões, identificando as variáveis mais relevantes para a geração de uma determinada saída.

Interpretar um modelo de aprendizagem e identificar as variáveis que mais ou menos contribuíram para uma determinada decisão são práticas que vem crescendo com o advento da aprendizagem de máquina no cenário atual. Por exemplo, um classificador que identifica uma determinada doença somente pode não ser interessante, mas quando acompanhado de uma interpretação que apresenta quais variáveis levaram o modelo a tomar tal decisão, se torna mais confiável, transparente e de maior aplicabilidade. Revisões acerca de modelos interpretáveis de aprendizagem podem ser vistos em Mi, Li e Zhou (2020) Agarwal e Das (2020).

Dentre os modelos de interpretação apresentados nas revisões mencionadas acima, dois tipos podem ser levantados: modelos de interpretação globais, que interpretam conjuntos de dados, mostrando as variáveis mais relevantes para todas as amostras incluídas naquele determinado conjunto e modelos locais, que avaliam o peso de cada variável para a geração da saída para uma única amostra. Inclusos no primeiro título, podem ser listados o *Partial Dependence Plot* (FRIEDMAN, 2001) e o *Permutation Feature Importance* (FISHER; RUDIN; DOMINICI, 2019) e o SHAP (acrônimo em inglês de *SHapley Additive exPlanations*) (LUNDBERG; LEE, 2017). Dentre os modelos locais, pode ser citado o LIME (sigla em inglês de *Local Interpretable Model-agnostic Explanations*) (RIBEIRO; SINGH; GUESTRIN, 2016) e o SHAP em algumas de suas análises.

Do ponto de vista da aplicação, a área da saúde tem realizado boas aplicações de modelos interpretáveis, uma vez que, descobrir as variáveis mais relevantes de um processo são cruciais no auxílio de tomada de

decisões. Em Karatekin et al. (2019), os autores usam a interpretabilidade para investigar os fatores de maior impacto na causa da retinopatia da prematuridade. Ramchandani, Fan e Mostafavi (2020) propuseram um modelo de *deep-learning* interpretável para investigar o crescimento das infecções por Covid-19.

Outra aplicação pode ser vista em Azodi, Tang e Shiu (2020) onde os autores investigam a interpretabilidade em aprendizagem de máquina com enfoque na genética. Além dos modelos de interpretação que fazem uso de um modelo de aprendizagem já treinado, existem modelos que, por si só, já possuem interpretabilidade, como o caso de sistemas Fuzzy, onde é possível construir ou extrair regras para o sistema inteligente, permitindo, assim, a interpretação por meio de termos linguísticos (VUČETIĆ; HUDEC; BOŽILOVIĆ, 2020). Além dos sistemas Fuzzy, já conhecidos, outros modelos com interpretabilidade embutida nos mesmos (sem a necessidade de um modelo interpretador) vêm sendo estudados e desenvolvidos (HOU; ZHOU, 2020).

2.5.1 LIME - *Local Interpretable Model-agnostic Explanations*

O LIME (RIBEIRO; SINGH; GUESTRIN, 2016) é um modelo de interpretação local de algoritmos de aprendizagem de máquina do tipo *Model-agnostic*, ou seja, não necessita de informações do modelo de aprendizagem desenvolvido, apenas da saída do mesmo. O LIME tem como o princípio o uso de um modelo preditivo local de baixa complexidade e boa interpretabilidade para explicar modelos de aprendizagem mais complexos. A formulação matemática do LIME tem como objetivo minimizar a diferença entre a predição do modelo previamente desenvolvido e o modelo local de interpretação, dada pela equação 2.13:

$$\xi = \arg\min_{g \in G} \mathscr{L}(f, g, \pi_x) + \Omega(g) \ , \tag{2.13}$$

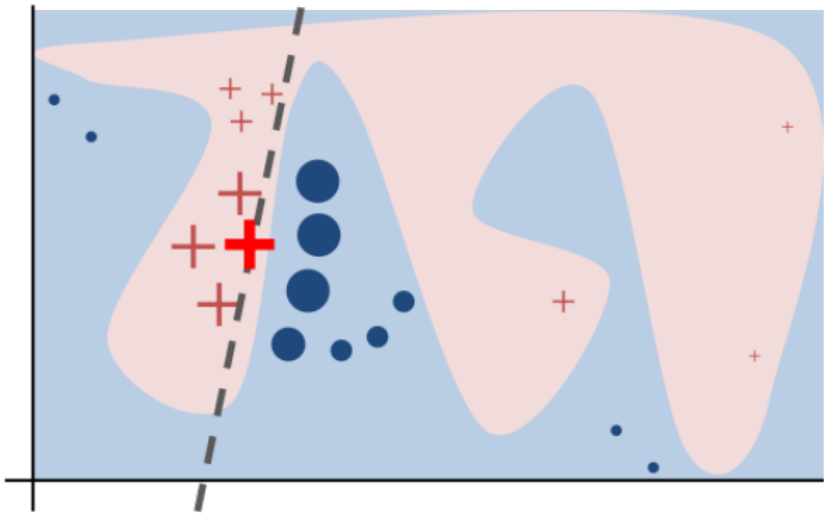
onde ξ é a diferença entre as predições, $f(x)$ função do modelo treinado, g a função do modelo aproximador local, G a saída do modelo aproximador, $\Omega(g)$ a função de complexidade do modelo treinado e π_x proximidade da instância x à vizinhança criada em torno desta e $\mathscr{L}$ a função de perda local quadrática.

O princípio de funcionamento do LIME consiste na obtenção da saída de uma dada entrada fornecida e, a partir deste ponto de saída, são geradas perturbações ao entorno deste ponto. Estas perturbações são geradas de forma aleatória como, por exemplo, ruídos gaussianos (MOLNAR, 2020). A partir deste pequeno conjunto de dados composto pela entrada, a saída predita e suas respectivas perturbações, é ajustado um modelo local simplificado.

Como modelos locais simplificados, podem ser utilizados tanto as Árvores de Decisão quanto um modelo de Regressão Linear. Ambos possuem boa interpretabilidade, seja por meio das regras geradas pela

árvore ou pelos coeficientes (pesos) dos regressores. Um exemplo do funcionamento do LIME pode ser visto na Figura 2.8.

Figura 2.8 – Exemplo do funcionamento do LIME em um problema de classificação com fronteira de decisão complexa. Os ponto em cruz destacado é a instância a ser interpretada, os demais pontos em cruz são perturbações de uma classe e os pontos em círculo ao redor são perturbações da outra classe também gerados pelo LIME. A linha tracejada refere-se ao modelo local gerado.



Fonte: Ribeiro, Singh e Guestrin (2016).

Após a obtenção do modelo local, o LIME realiza a interpretação por meio de gráfico ou tabela, indicando o intervalo de validade de cada variável para a interpretação e o valor do peso de cada variável e seus respectivos sinais. O sinal do peso indica se esta variável contribui para o aumento ou redução da saída para um problema de regressão. Em problemas de classificação, o sinal do peso indica se contribui para o aumento ou redução da probabilidade da observação pertencer à determinada classe.

Ribeiro, Singh e Guestrin (2016) também propuseram uma extensão do LIME com o objetivo de interpretar um determinado conjunto de amostras, o *Submodular Pick* (SP-LIME). O SP-LIME obtém a interpretação global por meio da geração de várias interpretações locais em conjunto com um algoritmo de otimização. Esta função busca maximizar a cobertura dos componentes mais importantes para a interpretação do modelo de aprendizagem. A função de otimização do *Submodular Pick* é determinada pela equação 2.14

$$Pick(W,I) = \arg\max_{V,|V|\leq B} c(V,W,I) \tag{2.14}$$

onde c é a função de cobertura das importâncias, B o número de explicações que o usuário deseja, V o conjunto de dados, W corresponde à matriz de explicação e I computa a importância total das variáveis.

Nas Figuras 2.9, 2.10 e 2.11 são apresentados um exemplo de saída do LIME em formato de tabela, formato gráfico para uma instância e uma saída do *Submodular Pick*. Nas figuras apresentadas, podem ser vistos os intervalos para os quais a saída é válida fornecendo uma forma de leitura da interpretação fornecida pelo algoritmo.

As interpretações obtidas pelo LIME para uma única amostra podem ser analisadas da seguinte forma: baseando-se na Figura 2.10, observa-se que o atributo 'flavonoids' quando maior ou igual à 1.04, contribuiu para um aumento da probabilidade desta amostra pertencer à classe 2. Enquanto o atributo 'color intensity', quando situado entre 4.60 e 6.15 contribui da mesma forma que o 'flavonoids', entretanto, em menor intensidade. De forma contrária, o atributo 'hue', entre os valores 0.77 e 0.97 contribui para a redução da probabilidade desta instância pertencer à classe 2.

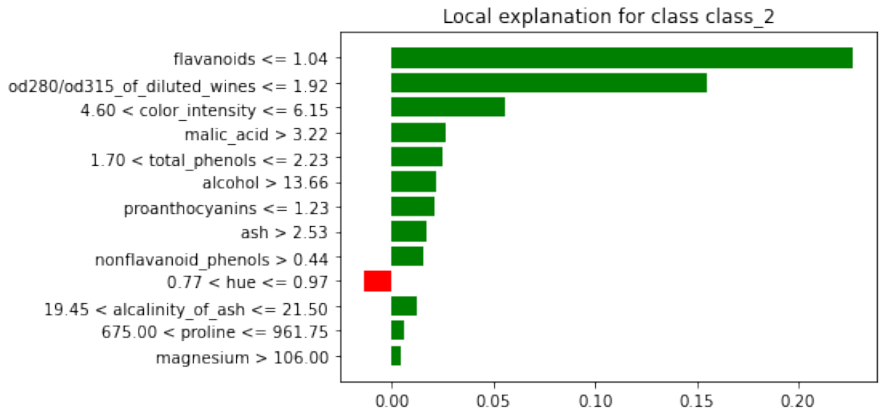
Para o SP-LIME, as interpretações operam para todo o conjunto de dados ao qual o modelo foi aplicado. Desta forma, um exemplo de interpretação segue para o atributo 'alcohol', o qual, quando menor que 13.66, implica em um aumento, em maior intensidade, da probabilidade do conjunto de dados pertencer à classe 0, contribui para uma redução da probabilidade das instâncias pertencerem à classe 1 e contribui para um aumento da probabilidade das amostras pertencerem à classe 2, porém, em menor intensidade que o aumento para a classe 0.

Figura 2.9 – Exemplo de saída do LIME em formato de Tabela.

0	flavanoids <= 1.04	0.226535
1	od280/od315_of_diluted_wines <= 1.92	0.155038
2	4.60 < color_intensity <= 6.15	0.055943
3	malic_acid > 3.22	0.026609
4	1.70 < total_phenols <= 2.23	0.024925
5	alcohol > 13.66	0.022163
6	proanthocyanins <= 1.23	0.021203
7	ash > 2.53	0.017328
8	nonflavanoid_phenols > 0.44	0.015858
9	0.77 < hue <= 0.97	-0.013321
10	19.45 < alcalinity_of_ash <= 21.50	0.012182
11	675.00 < proline <= 961.75	0.006224
12	magnesium > 106.00	0.004680

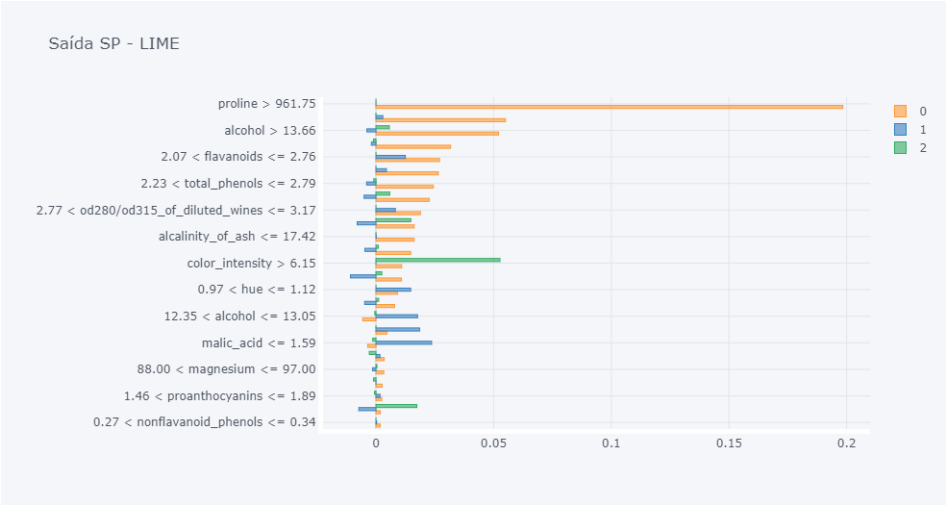
Fonte: Do Autor (2021).

Figura 2.10 – Exemplo de saída gráfica do LIME para uma amostra. O eixo x representa o peso da interpretação.



Fonte: Do Autor (2021).

Figura 2.11 – Exemplo de saída gráfica do SP-LIME. O eixo x representa o peso da interpretação.



Fonte: Do Autor (2021).

3 MATERIAIS E MÉTODOS

Para a execução deste trabalho, foram utilizadas ferramentas de *software* e uma infraestrutura de *hardware* que possibilite a execução da aplicação proposta. Como ferramentas de *software*, a Linguagem *Python*, versão 3.9 foi escolhida para a análise de dados e aplicação dos algoritmos de classificação. A escolha pelo uso da linguagem Python se deu pela sua facilidade de uso, vasta disponibilidade de bibliotecas voltadas para ciência de dados e aprendizagem de máquina, comunidade ativa em fóruns de discussão para resolução de dúvidas e solução de *bugs*, além de ser uma linguagem livre, o que não requer uma licença proprietária. As principais bibliotecas utilizadas neste trabalho foram: o *numpy*[1] para cálculos matemáticos em geral; *pandas*[2] que opera em conjuntos de dados no formato de planilhas, auxilia na organização de *datasets* e em levantamentos descritivos da base de dados; *scikit-learn*[3] para a implementação dos algoritmos classificadores; e o *matplotlib*[4] para a geração de gráficos.

Como infraestrutura de *hardware*, foram utilizados 2 computadores com configurações diferentes: o primeiro constituído de um processador Intel Core i5-1135G7, com frequência de base de 2,40GHz, memória RAM de 8GB DRR4 com frequência de 2666MHz. O segundo computador, pertencente ao Laboratório de Processamento de Dados da Universidade Federal de Lavras, possui processador Intel Core i7-7700, com frequência de base de 3,60GHz e memória RAM de 16GB DDR4 com frequência de 2400MHz.

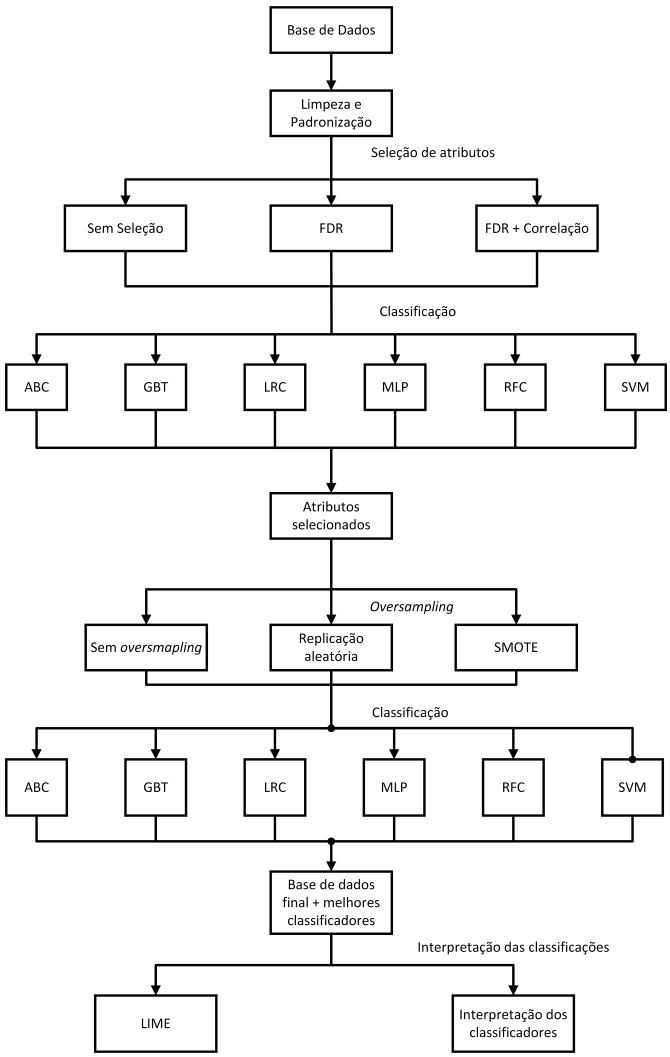
Para ilustrar o fluxo de atividades do sistema proposto, de maneira a obter uma visão simplificada, foi gerado um diagrama que apresenta cada etapa do trabalho realizado. O diagrama é mostrado na Figura 3.1 e mostra cada etapa do processamento dos registros do sistema de classificação de registros do CAR.

3.1 Base de Dados

A base de dados utilizada consiste em um compilado de atributos existentes no CAR composto por variáveis relacionadas ao imóvel rural, tais como: área do imóvel, número de módulos fiscais do terreno, vértices do polígono do terreno (dado como entrada em mapa desenhado na plataforma SiCAR ou por fornecimento de arquivo de georreferenciamento da propriedade); área das feições do imóvel rural, como rio, nascente, vegetações nativas (resinga, manguezal, vereda, etc.); e respostas de um questionário sobre a

[1] <https://numpy.org/>
[2] <https://pandas.pydata.org/>
[3] <https://scikit-learn.org/stable/>
[4] <https://matplotlib.org/>

Figura 3.1 – Diagrama representando o *setup* experimental.



Fonte: Do Autor (2022).

propriedade, onde o proprietário responde acerca do imóvel perguntas sobre a regularização ambiental do imóvel rural por meio de respostas objetivas (não, sim, não informar).

A base dados possui um total de 90 atributos que foram divididos em 6 grupos, sendo estes: Área do imóvel rural, Vértices do polígono do terreno, Retificações, Sobreposição, Questionário e Feições do imóvel rural. Na Tabela 3.1 são apresentados os atributos da base de dados em cada um de seus respectivos grupos. Detalhes explicativos sobre os atributos dispostos na base de dados podem ser vistos na cartilha de campanha do CAR, disponível pelo Serviço Florestal Brasileiro (Serviço Florestal Brasileiro, 2022a).

Tabela 3.1 – Atributos da base de dados em seus respectivos grupos

Grupo	Atributos
Área do imóvel rural	‘AREA_HA’; ‘ NUMERO_MF’; ‘AREA_DOC’
Vértices do polígono do terreno	‘ N_VERT_PER’
Retificações	‘ QTD_RETIFICACOES’
Sobreposição	‘ QTD_SOB_IR’; ‘SOBR_TI_HA’; ‘PERC_SO_TI’; ‘SOBR_UC_HA’; ‘PERC_SO_UC’
Questionário	‘ TAMANHO_ALTERADO_APOS_2008’; ‘POSSUI_CRF’; ‘EXISTE_RPPN’; ‘POSSUI_EXCEDENTE_VEGETACAO_NATIVA’; ‘EXISTE_INFRACAO’; ‘EXISTE_PRAD’; ‘EXISTE_TAC’; ‘POSSUI_DEFICIT_RL’; ‘DESEJA_ADERIR_PRA’
Feições do imóvel rural	‘AREA_CONSOLIDADA’; ‘VEGETACAO_NATIVA’; ‘AREA_POUSIO’; ‘AREA_INFRAESTRUTURA_PUBLICA’; ‘AREA_UTILIDADE_PUBLICA’; ‘RESERVATORIO_ENERGIA’; AREA_USO_RESTRITO_DECLIVIDADE_25_A_45’; ‘AREA_USO_RESTRITO_PANTANEIRA’; ‘RIO_ATE_10’; ‘RIO_10_A_50’; ‘RIO_50_A_200’; ‘RIO_200_A_600’; ‘RIO_ACIMA_600’; ‘LAGO_NATURAL’; ‘NASCENTE_OLHO_DAGUA’;‘RESERVATORIO_ARTIFICIAL_DECORRENTE_BARRAMENTO; ‘MANGUEZAL’; ‘VEREDA’; ‘AREA_ALTITUDE_SUPERIOR_1800’; ‘AREA_DECLIVIDADE_MAIOR_45’; ‘BORDA_CHAPADA’; ‘AREA_TOPO_MORRO’; ‘ARL_PROPOSTA’; ‘ARL_AVERBADA’; ‘ARL_APROVADA_NAO_AVERBADA’; ‘AREA_IMOVEL’; ‘AREA_IMOVEL_LIQUIDA’; ‘AREA_ENTORNO_RESERVATORIO_ENERGIA’; ‘AREA_SERVIDAO_ADMINISTRATIVA_TOTAL’; ‘APP_TOTAL’;‘APP_ESCADINHA’; ‘ARL_TOTAL’; ‘ARL_A_RECUPERAR’; ‘APP_A_RECUPERAR’; ‘RESTINGA’;‘APP_ESCADINHA_RIO_ATE_10’; ‘APP_ESCADINHA_RIO_10_A_50’; ‘APP_ESCADINHA_RIO_50_A_200’;‘APP_ESCADINHA_RIO_200_A_600’; ‘APP_ESCADINHA_RIO_ACIMA_600’; ‘APP_ESCADINHA_VEREDA’; ‘APP_ESCADINHA_LAGO_NATURAL’; ‘APP_ESCADINHA_NASCENTE_OLHO_DAGUA’; ‘APP_RIO_ATE_10’; ‘APP_RIO_10_A_50’; ‘APP_RIO_50_A_200’; ‘APP_RIO_200_A_600’; ‘APP_RIO_ACIMA_600’; ‘APP_NASCENTE_OLHO_DAGUA’; ‘APP_LAGO_NATURAL’; ‘APP_RESERVATORIO_ARTIFICIAL_DECORRENTE_BARRAMENTO’; ‘APP_VEREDA’; ‘APP_AREA_TOPO_MORRO’; ‘APP_MANGUEZAL’; ‘APP_AREA_ALTITUDE_SUPERIOR_1800’; ‘APP_BORDA_CHAPADA’; ‘APP_RESTINGA’; ‘APP_AREA_DECLIVIDADE_MAIOR_45’; ‘AREA_NAO_CLASSIFICADA’; ‘CORPO_DAGUA’; ‘APP_AREA_VN’; ‘APP_VAZIO’; ‘BANHADO’; ‘APP_BANHADO’; ‘AREA_IMOVEL_LIQUIDA_ANALISE’; ‘ARL_AVERBADA_OUTRO_IMOVEL’; ‘SEDE_IMOVEL’; ‘RESERVATORIO_GERACAO_ENERGIA_ATE_24_08_2001’; ‘APP_RESERVATORIO_GERACAO_ENERGIA_ATE_24_08_2001’; ‘AREA_TERRITORIO_PCT’

Fonte: Do Autor (2022).

Os registros fornecidos consistem na base de dados pública do CAR com a adição de cruzamentos espaciais para verificação de sobreposições espaciais entre os polígonos dos imóveis cadastrados. A base preparada foi fornecida pela Agência Zetta de Inovação pertencente à Universidade Federal de Lavras (UFLA), que desenvolve projetos relacionados ao sistema do CAR. Foram disponibilizados registros do CAR de 18 estados brasileiros, sendo estes: Acre, Alagoas, Amapá, Ceará, Espírito Santo, Goiás, Maranhão, Pará, Paraíba, Pernambuco, Piauí, Rio de Janeiro, Rio Grande do Norte, Rondônia, Roraima, São Paulo, Sergipe e Tocantins.

A base de dados possui como atributo de saída a condição do cadastro, que refere-se ao *status* do mesmo, possuindo 3 classes: aprovado sem pendências, cancelado ou constando pendências. Os números referentes aos registros por estado e condição estão contidos na Tabela 3.2. Dado que este trabalho optou por uma abordagem supervisionada de classificação em cadastros a serem aprovados ou cancelados, foram utilizados somente os registros rotulados, que passaram pelas análises manuais e tiveram a tomada de decisão

realizada. Esta abordagem foi escolhida devido aos registros pendentes necessitarem de uma análise para sua rotulação adequada, ou seja, os cadastros pendentes possuem rotulação indeterminada.

Portanto, os registros pendentes não foram utilizados nos ensaios de classificação, todavia, dado que o número de registros pendentes é elevado, uma análise estatística descritiva por meio da biblioteca disponível na linguagem *Python Pandas Profiling*[5], foi realizada. A análise faz um *profile* estatístico[6] dos atributos da base de dados, podendo, assim, observar o comportamento dos registros pendentes e encontrar possíveis similaridades entre a base rotulada e não rotulada, gerando possíveis apontamentos futuros para outras abordagens utilizando os registros pendentes, que consistem na maior parcela dos cadastros do CAR. Logo, como variável de saída (classe), têm-se 2 valores: 'cancelado' (0) ou 'aprovado' (1).

Tabela 3.2 – Números da base de dados utilizada para os experimentos

UF	Aprovados	Cancelados	Pendentes	Total
AC	14	322	38505	38841
AL	0	100	97412	97512
AP	0	258	7173	7431
CE	1456	2535	266313	270304
ES	0	540	96524	97064
GO	40	578	176402	177020
MA	20	781	190809	191610
PA	159	3819	0	3978
PB	51	280	148376	148707
PE	0	69	284460	284529
PI	0	162	215486	215648
RJ	160	400	53265	53825
RN	0	138	76508	76646
RO	323	1571	118058	119952
RR	0	19	12776	12795
SE	2	87	78827	78916
SP	632	4983	379260	384875
TO	3454	9505	70003	82962
Total	6311	26147	2310157	2342615

Fonte: Do Autor (2022).

[5] <https://pandas-profiling.ydata.ai/docs/master/index.html>

[6] Trata-se de uma análise exploratória da base de dados, contendo histogramas de cada atributo, correlações, métricas estatísticas (média, variância), dentre outras análises.

3.2 Pré - Processamento

Para padronizar o conjunto de dados obtido, foram necessárias algumas adaptações para que o conjunto de dados pudesse ser tratado por um algoritmo de aprendizagem de máquina. Primeiramente, foram removidas algumas variáveis que não possuíam informações pertinentes, tais como 'UF do imóvel', 'município', 'id do imóvel', 'tipo de documento da área inserido' e informações sensíveis, como o nome do cadastrante.

Após a remoção dos atributos considerados não pertinentes, foi realizada uma codificação do questionário para as perguntas objetivas, tal que para as respostas 'sim', 'não' e 'não informado', foram atribuídos os respectivos valores 1; -1 e 0. Outra adaptação de variáveis realizada foi a soma das áreas dos terrenos fornecidos (para o caso dos cadastros com mais de uma área inserida). Utilizando o *profile* estatístico, foi constatada a presença de atributos constantes em zero, logo, estas variáveis também foram removidas, pois não possuem relevância para a discriminação entre as classes.

Seguinte à etapa da limpeza e codificação dos atributos, foi realizado o procedimento de seleção de atributos, por meio do FDR que gera o vetor de pesos que indica a relevância de cada atributo na classificação. O vetor de pesos foi ordenado de forma decrescente e o número de *features* a serem utilizadas de acordo com a relevância apresentada pelo FDR foi variado da seguinte forma: primeiramente, entre 10 e 70 atributos, variando de 5 em 5. Em seguida, o intervalo de melhor desempenho (ex.: intervalo entre 45 e 50 atributos) que obteve o melhor desempenho foi explorado com variação unitária até o valor de *features* que obteve o melhor desempenho preditivo (acurácia). Para a normalização dos atributos, foi utilizada a normalização *Z-Score*, definida pela equação 3.1:

$$\widetilde{X}_i = \left(\frac{X_i - \mu(x_i)}{\sigma(x_i)} \right) \tag{3.1}$$

onde $\widetilde{X}_i$ é o vetor da i-ésima variável da base de dados normalizado, $\mu(x_i)$ e $\sigma(x_i)$, a média e o desvio-padrão de X_i, respectivamente.

Realizada a seleção de atributos, foram divididos em dois conjuntos: o de treinamento e teste com validação cruzada e o conjunto de validação final, contido por registros novos aos classificadores. A proporção desta divisão foi de 80% para o conjunto de treinamento e teste e os 20% restantes foram utilizados para validação final. Cabe salientar que esta divisão respeitou a distribuição entre as classes, portanto a proporção foi em relação ao tamanho amostral de cada classe individualmente. Em seguida, foram aplicadas as técnicas

de *oversampling* para o balanceamento entre as classes para o conjunto de treinamento. Desta maneira, o conjunto de validação será contido somente por dados reais, evitando assim, eventuais vieses causados pela super-amostragem. Os tamanhos dos conjuntos contendo somente dados reais para treinamento e teste em cada classe estão dispostos na Tabela 3.3.

Tabela 3.3 – Conjuntos de dados de treinamento e teste

Classe	Treinamento + Teste (sem *oversampling*)	Treinamento + Teste (com *oversampling*)	Validação	Total
Aprovados - 1	5048	20917	1263	6311
Cancelados - 0	20917	20917	5230	26147
Total	25965	41834	6493	32458

Fonte: Do Autor (2022).

3.3 Avaliação dos modelos e uso do interpretador

A partir desta seção, visando uma simplificação, será feita uma abreviação dos nomes dos classificadores mencionados na seção 3.3 da seguinte forma: ABC (*AdaBoost Classifier*), GBT (*Gradient Boosting Tree*), LRC (*Logistic Regression Classifier*), MLP (*Multi-Layer Perceptron*), RFC (*Random Forest Classifier*) e SVM (*Support Vector Machines*).

Realizados os procedimentos de pré-processamento, foram realizados testes de avaliação dos modelos e ensaios comparativos avaliando cada etapa de projeto do método. Os hiperparâmetros dos classificadores foram ajustados de acordo com variações manuais dentro de uma determinada faixa de valores sob validação cruzada do tipo *k-fold* com 10 *folds*. As faixas dos hiperparâmetros utilizados para cada classificador estão dispostas na Tabela 3.4.

Dos parâmetros mencionados na Tabela 3.4, o único que possuiu alteração de acordo com o método de seleção de atributos foi o número de neurônios na camada oculta para o classificador MLP, cuja variação nos número de neurônios nas camadas ocultas foi de acordo com o número de *features* da base de dados. Esta escolha foi realizada, uma vez que, durante os testes experimentais, o aumento no número de neurônios não implicou diretamente em um aumento do desempenho preditivo. Portanto, como os valores do hiperparâmetro foram alterados em cada ensaio, foi escolhido alocar a variação entre os neurônios em uma tabela a parte (vide Tabela 3.5).

O critério de ajuste dos hiperparâmetros dos classificadores se deu pela medida da acurácia, portanto, o classificador que será utilizado para a validação final será aquele que obtiver a melhor acurácia durante o

Tabela 3.4 – Faixa de valores dos hiperparâmetros utilizados nos experimentos

Modelo	Hiperparâmetro	Faixa de valores
ABC	Medida de avaliação	[‘gini’, ‘entropia’]
	Estratégia de divisão de folhas	[‘melhor’, ‘aleatória’]
	Profundidade máxima	[3, 5, 7, 9, 11, 13, 15]
	Número de árvores	[20, 30, 50, 80, 100, 120, 150, 200, 250, 300, 400, 500]
GBT	Profundidade máxima	[20, 30, 50, 80, 100, 120, 150, 200, 250, 300]
	Taxa de aprendizado	[0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2]
	Número de árvores	[3, 5, 7, 9, 11, 13, 15, 17, 19]
LRC	C	[0.0001, 0.0003, 0.0007, 0.0018, 0.0048, 0.0127, 0.0336, 0.0886, 0.2336, 0.6158, 162378, 428133, 112884, 297635, 78476, 206914, 545559, 1438.45, 3792.69, 10000]
	Solver	[‘newton-cg’, ‘lbfgs’, ‘sag’]
	Regularização	[‘nenhuma’, ‘l1’, ‘l2’, ‘elasticnet’]
MLP	ativacao	[‘tangente hiperbólica’, ‘relu’]
	Solver	[‘sgd’, ‘adam’, ‘lbfgs’]
	Taxa de aprendizado	[‘constante’,‘adaptativo’]
RFC	Profundidade máxima	[3, 5, 7, 9, 11, 13, 15, 17, 19]
	Medida de avaliação	[‘gini’, ‘entropia’]
	Número de árvores	[20, 30, 50, 80, 100, 120, 150, 200, 250, 300, 400, 500]
SVM	C	[0.1, 1, 10, 50, 100, 500, 1000]
	γ	[‘scale’, ‘automático’]

Fonte: Do Autor (2022).

Tabela 3.5 – Faixa de valores dos neurônios nas camadas ocultas para cada método de seleção de atributos

Método de seleção de atributos	Número de Neurônios na camada oculta
Sem seleção de atributos	(10), (15), (20), (25), (30), (35), (40), (45), (50), (55), (60), (65), (70), (75), (80), (10,10), (20,10), (40,10), (60,10), (80,10)]
FDR	(10), (15), (20), (25), (30), (35), (40), (45), (50), (10,10), (10,15), (10,20), (10,30), (10,40), (10,50)]
FDR + Correlação	(10), (15), (20), (25), (30), (35), (40), (45), (10,10), (10,15), (10,20), (10,25), (10,30), (10,40)]

Fonte: Do Autor (2022).

processo de variação dos hiperparâmetros no treinamento com o *k-fold*. Encontrado os melhores hiperparâmetros, os classificadores serão retreinados com toda a base de dados de treinamento para um ajuste fino dos parâmetros, uma vez que o classificador foi treinado com 90% do conjunto de treinamento, tendo os outros 10% utilizados para teste no processo do *k-fold*. Realizado o processo de retreinamento dos classificadores, os mesmos serão avaliados com o conjunto de dados de validação, este contendo registros não vistos previamente pelos modelos. Para os ensaios comparativos, foram gerados resultados numéricos e gráficos para os conjuntos de treinamento, teste e validação. Como medidas numéricas, foram utilizadas a acurácia do modelo (ACC), o *F1-Score*, a AUC e as medidas de *precision* e *recall* para cada classe e como resultados gráficos foram geradas as curvas ROC para cada modelo e mostradas para comparativos.

Além das medidas de avaliação, será comparada a acurácia de cada modelo durante a etapa de treinamento e teste com validação cruzada por meio do teste estatístico *t* com validação cruzada do tipo 5x2. Este teste estatístico verifica se o desempenho de um classificador possui diferença estatística significativa em relação a outro por meio do teste de hipótese t de *Student*. Por meio deste teste, caso o valor *p* seja maior que o valor de significância estatístico definido, aceita-se a hipótese nula, dizendo que os dois classificadores não possuem diferença estatística significativa, caso contrário, rejeita-se a hipótese nula e assume-se que os desempenhos dos classificadores possuem diferença estatística significativa. O tipo de teste estatístico realizado será o do tipo 5x2 com validação cruzada (DIETTERICH, 1998). O teste funciona da seguinte maneira: São realizados 5 treinamentos dos modelos sob validação cruzada *k-fold* com 2 *folds*, assim, são tomadas as acurácias do conjunto de teste e realizado o teste de hipótese.

O teste será realizado para cada par de classificadores, onde será verificado se o desempenho entre estes possui significância estatística. Para este teste, será adotado o nível de significância de 5%, ou seja, caso o valor de *p* obtido nos testes seja maior que 0.05, será aceita a hipótese nula, do contrário, a hipótese nula será rejeitada e será assumida a diferença estatística significativa entre o par de classificadores. Os testes t serão realizados utilizando a API *mlxtend*[7] disponível na linguagem *Python*.

O fluxo dos ensaios comparativos seguirá da seguinte forma: serão comparados, primeiramente, os algoritmos de seleção de atributos, analisando os modelos treinados sem seleção de *features* (apenas a limpeza inicial dos atributos sem relevância), os modelos treinados sob a seleção de atributos realizada pelo FDR. Esta primeira etapa tem por objetivo escolher qual o método de seleção foi o mais eficiente, baseando-se em 2 critérios: o número de atributos selecionados e o desempenho preditivo de classificação.

Seguinte ao comparativo realizado com a seleção de atributos, foi realizado o teste entre os classificadores treinados em relação ao tipo de *oversampling*. Como *baseline* para os testes comparativos, foi utilizado o conjunto de dados com o melhor desempenho no processo de seleção de atributos. Serão comparados os seus resultados preditivos com a base de dados desbalanceada juntamente com a base de dados balanceada pela replicação aleatória e pelo SMOTE. Conforme a reamostragem aleatória não possui hiperparâmetros para ajuste, esta será aplicada diretamente, enquanto o SMOTE terá o hiperparâmetro k ajustando entre $k = [3,5,7,9,11,13,15]$, sendo escolhido o valor que obtiver a melhor acurácia. Este segundo comparativo tem por finalidade avaliar qual o melhor classificador sob qual forma de conjunto de dados de treinamento. O objetivo é verificar se o treinamento com o conjunto de dados original representa um melhor desempenho ou se algum dos métodos de superamostragem auxiliam no aumento do desempenho dos classificadores.

[7] <http://rasbt.github.io/mlxtend/user_guide/evaluate/paired_ttest_5x2cv/>

Finalizados os ensaios comparativos entre os classificadores, segue o teste de interpretação das classificações obtidas, esta foi realizada da seguinte forma: primeiramente, o algoritmo de melhor desempenho geral foi aplicado ao LIME para que o mesmo gere as interpretações. Para isto, foi utilizada a extensão *Submodular Pick* para realizar uma análise geral do conjunto de dados de teste de forma a analisar quais variáveis tem maior impacto em cada classe, além do intervalo no qual estas possuem este impacto na saída.

Além das interpretações do LIME, foram utilizadas as interpretações internas de classificadores que permitem tal propriedade, estes sendo a Regressão Logística, a *Random Forest* e o *Gradient Boosting*. Tais ensaios comparativos buscam verificar quais dos algoritmos podem ser melhor aplicados para uma interpretação da classificação, podendo corroborar com as avaliações realizadas pelas análises manuais ou se as interpretações apresentadas não possuem correspondência prática. O objetivo final desta análise é elencar quais variáveis são mais influentes na tomada de decisão do modelo de aprendizagem, permitindo a observação de quais mais influenciam na aprovação ou reprovação do cadastro e como influenciam.

4 RESULTADOS E DISCUSSÃO

Neste capítulo serão apresentados e discutidos os resultados obtidos neste trabalho, divididos pelas seguintes seções. Na seção 4.1 serão apresentados os resultados da análise exploratória dos registros do CAR. Os resultados de classificação comparando os métodos de seleção de atributos são apresentados e discutidos na seção 4.2 e os resultados e discussão acerca da classificação comparando os métodos de *oversampling* estão inseridos na seção 4.3. A apresentação dos resultados de inerpretação das classificações obtidas e sua discussão estão inseridas na seção 4.4.

4.1 Resultados exploratórios com o *profile* estatístico

Esta análise inicial foi realizada com duas principais finalidades: (i) fazer um levantamento das estatísticas descritivas dos atributos da base de dados com os registros rotulados e os registros pendentes; e (ii) verificar a possibilidade de um futuro uso dos registros pendentes para trabalhos de classificação como, por exemplo, em uma abordagem de classificação semi-supervisionada. Assim, foram levantadas as medidas estatísticas de média, desvio-padrão, variância, curtose, assimetria e o valor do percentil em 95%. Dado que o número de atributos é grande para alocar em uma única tabela, foram geradas 3 tabelas para as estatísticas descritivas (Tabelas 4.1, 4.2 e 4.3).

Tabela 4.1 – Estatísticas descritivas para os atributos da base de dados de cadastros rotulados e para a base de dados de cadastros pendentes - Parte 1

Atributo	Média		Desvio-padrão		Variância		Curtose		Assimetria		Percentil 95%	
	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes
AREA_HA	6237984,55	277208,81	60049618,03	3600226,39	3,61E+15	1,30E+13	4189,47	383284,44	58,88	476,39	1,51E+07	890385,60
NUMERO_MF	79727,62	6750,61	496394,79	79291,88	2,46E+11	6,29E+09	2635,42	60848,91	43,27	180,12	250007,00	26072,20
N_VERT_PER	79,65	23,02	253,40	43,72	64210,88	1911,5	3926,72	107,32	46,69	8,50	614,00	74,00
QTD_RETIFICACOES	0,54	0,33	1,12	4,48	1,26	20,08	36,85	134145,63	3,87	359,65	3,00	2,00
QTD_SOB_IR	1,51	2,95	5,32	4,80	28,26	23,01	3243,19	1370,30	43,48	30,30	6,00	7,00
SOBR_TI_HA	841,87	364,26	66643,48	129656,04	4,44E+09	1,68E+10	11632,27	1476446,38	105,81	1132,91	0,00	0,00
PERC_SO_TI	588,43	236,68	21969,53	14835,85	4,83E+08	2,20E+08	1825,60	4355,14	42,15	65,62	0,00	0,00
SOBR_UC_HA	2466357,03	131470,51	21281725,23	2431265,69	4,53E+14	5,91E+12	8437,59	310935,15	71,66	388,10	3,64E+06	76,10
PERC_SO_UC	94114,61	46537,63	314866,87	223755,30	9,91E+10	5,01E+10	12,06	28,98	3,45	5,17	1,00E+06	100,01
AREA_DOC	1244,25	1698,19	74515,20	609571,48	5,55E+09	3,72E+11	24442,77	1179652,10	154,41	972,71	1583,25	230,00
TAMANHO_ALTERADO_APOS_2008	-0,26	-0,68	0,53	0,55	0,28	0,30	-0,41	1,22	-0,13	1,48	0,00	0,00
POSSUI_CRF	-0,32	-0,74	0,47	0,45	0,22	0,20	-1,31	0,24	-0,68	1,31	0,00	0,00
EXISTE_RPPN	-0,33	-0,76	0,47	0,43	0,22	0,18	-1,44	-0,55	-0,73	1,20	0,00	0,00
POSSUI_EXCEDENTE_VEGETACAO_NATIVA	0,16	-0,29	0,80	0,83	0,65	0,69	-1,40	-1,31	-0,30	0,57	1,00	1,00
EXISTE_INFRACAO	-0,32	-0,73	0,48	0,48	0,23	0,23	-1,19	0,95	-0,64	1,43	0,00	0,00
EXISTE_PRAD	-0,32	-0,76	0,47	0,43	0,22	0,19	-1,40	-0,49	-0,72	1,21	0,00	0,00
EXISTE_TAC	-0,32	-0,76	0,47	0,43	0,22	0,19	-1,37	-0,47	-0,71	1,21	0,00	0,00
POSSUI_DEFICIT_RL	-0,20	-0,39	0,56	0,78	0,31	0,61	-0,21	-0,91	-0,02	0,80	1,00	1,00
DESEJA_ADERIR_PRA	0,30	0,36	0,62	0,83	0,39	0,68	-0,66	-1,13	-0,31	-0,75	1,00	1,00
AREA_CONSOLIDADA	179,72	27,42	1034,28	176,10	1069730,18	31012,92	3299,17	9561,57	43,86	63,93	582,51	97,77
VEGETACAO_NATIVA	297,08	30,69	1956,09	453,68	3826280,55	205822,83	12098,89	167579,13	94,44	308,66	1496,86	89,21
AREA_POUSIO	2,13	0,90	289,24	22,76	83658,97	517,98	25819,38	13769,96	160,47	79,89	0,00	0,00
AREA_INFRAESTRUTURA_PUBLICA	0,06	0,09	1,56	2,39	2,42	5,70	3242,55	115874,17	51,05	225,62	0,00	0,00
AREA_UTILIDADE_PUBLICA	0,06	0,06	3,50	7,31	12,27	53,48	22341,89	1211356,57	144,87	1003,52	0,00	0,02
RESERVATORIO_ENERGIA	0,05	0,05	5,49	12,42	30,12	154,34	22631,51	344616,41	147,11	547,98	0,00	0,00
AREA_USO_RESTRITO_DECLIVIDADE_25_A_45	0,61	0,38	19,75	27,77	389,94	770,99	9463,01	659661,67	84,23	709,66	0,00	0,00
AREA_USO_RESTRITO_PANTANEIRA	0,00	0,01	0,00	1,08	0,00	1,18	0,00	297361,12	0,00	460,79	0,00	0,00
RIO_ATE_10	4,82E+37	2,33E+52	7,77E+39	3,53E+55	6,04E+79	2310146,84	25965,00	2310146,84	161,14	1519,92	16,34	1,27
RIO_10_A_50	0,23	6,07E+44	5,08	6,32E+47	25,77542476	4,00E+95	5172,87	1612123,54	62,83	1233,36	0,00	0,00
RIO_50_A_200	6,54E+46	7,62E+42	1,05E+49	1,16E+46	1,11E+98	1,34E+92	25965,00	2310157,00	161,14	1519,92	0,00	0,00

Fonte: Do Autor (2022).

Tabela 4.2 – Estatísticas descritivas para os atributos da base de dados de cadastros rotulados e para a base de dados de cadastros pendentes - Parte 2

Atributo	Média		Desvio-padrão		Variância		Curtose		Assimetria		Percentil 95%	
	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes
RIO_200_A_600	0,64	0	33,02	0,00	1090,26	0,00	12352,14	0,00	102,53	0,00	0,00	0,00
RIO_ACIMA_600	1,42	0,07	59,90	7,61	3587,67	57,85	5272,16	141737,22	65,23	310,44	0,00	0,00
LAGO_NATURAL	2,14	8,01E+45	156,92	8,94E+48	24625,40	8,00E+97	23090,40	2026385,02	148,68	1393,36	5,18	0,00
NASCENTE_OLHO_DAGUA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
RESERVATORIO_ARTIFICIAL_DECORRENTE_BARRAMENTO	0,20	2,62E+53	6,94	3,98E+56	48,20	1,59E+113	18449,22	2310157,00	127,00	1519,92	0,00	0,26
MANGUEZAL	0,02	0,01	2,46	2,82	6,03	7,93	20103,59	212366,02	138,06	394,13	0,00	0,00
VEREDA	0,02	2,29E+45	1,34	3,33E+48	1,79	1,11E+97	10481,31	2304667,67	96,18	1517,27	0,00	0,00
AREA_ALTITUDE_SUPERIOR_1800	1,02E-05	0,00	0,00	0,52	1,47E-06	0,27	16513,16	1710914,20	125,78	1246,27	0,00	0,00
AREA_DECLIVIDADE_MAIOR_45	0,05	0,06	2,61	5,12	6,82	26,17	14772,71	350276,60	111,18	488,67	0,00	0,00
BORDA_CHAPADA	0,43	6,31E+43	16,24	9,20E+46	263,59	8,45E+93	4775,31	2303264,96	63,88	1516,60	0,00	0,00
AREA_TOPO_MORRO	0,05	0,06	2,35	4,02	5,52	16,18	7564,01	143341,40	76,89	276,79	0,00	0,00
ARL_PROPOSTA	205,78	17,96	1805,07	270,95	3258292,13	73414,80	16706,43	105682,81	117,27	233,52	615,09	50,98
ARL_AVERBADA	9,02	1,72	430,02	113,28	184917,12	12833,16	9609,39	745004,87	91,50	736,30	0,00	0,00
ARL_APROVADA_NAO_AVERBADA	0,87	0,18	57,57	15,18	3314,36	230,34	15649,23	107872,49	117,86	274,50	0,00	0,00
AREA_IMOVEL	886,84	79,70	11487,97	772,36	1,32E+08	596536,11	5685,54	123986,14	66,01	265,73	1958,18	257,31
AREA_IMOVEL_LIQUIDA	886,48	79,27	11487,80	738,41	1,32E+08	545254,67	5685,88	130554,48	66,02	264,06	1952,36	256,33
AREA_ENTORNO_RESERVATORIO_ENERGIA	0,01	0,01	0,78	2,17	0,61	4,72	23189,52	272736,58	149,56	451,38	0,00	0,00
AREA_SERVIDAO_ADMINISTRATIVA_TOTAL	0,37	0,27	11,98	16,23	143,40	263,27	7526,24	248715,20	79,96	454,15	0,00	0,38
APP_TOTAL	23,60	3,11	111,23	23,52	12371,76	553,11	1923,77	12783,45	35,32	72,23	95,67	11,58
APP_ESCADINHA	1,41	0,26	14,26	3,23	203,36	10,46	8457,22	26787,37	75,74	96,72	6,96	0,55
ARL_TOTAL	179,22	19,03	806,63	293,03	650652,87	85867,84	1103,94	93837,69	23,95	226,73	614,95	53,71
ARL_A_RECUPERAR	0,00	0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
APP_A_RECUPERAR	0,00	0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
RESTINGA	0,00	0,01	0,02	5,70	5,40E-04	32,45	9688,18	1557521,40	94,55	1165,51	0,00	0,00
APP_ESCADINHA_RIO_ATE_10	1,31	0,26	9,86	3,51	97,16	12,35	1102,70	15774,43	27,65	86,93	0,00	0,52
APP_ESCADINHA_RIO_10_A_50	0,03	0,01	1,37	0,49	1,87	0,24	10406,84	85073,50	93,69	221,10	0,00	0,00
APP_ESCADINHA_RIO_50_A_200	0,03	0,00	2,28	0,45	5,19	0,20	22343,52	305179,53	144,92	464,34	0,00	0,00
APP_ESCADINHA_RIO_200_A_600	0,05	0,00	5,74	0,35	32,93	0,12	23865,67	496981,82	151,84	599,60	0,00	0,00
APP_ESCADINHA_RIO_ACIMA_600	0,03	0,00	3,29	0,14	10,80	0,02	23374,11	298965,83	149,77	467,74	0,00	0,00
APP_ESCADINHA_VEREDA	4,95E-03	0,01	0,56	0,58	0,31	0,33	24732,83	133311,18	155,61	294,77	0,00	0,00

Fonte: Do Autor (2022).

Tabela 4.3 – Estatísticas descritivas para os atributos da base de dados de cadastros rotulados e para a base de dados de cadastros pendentes - Parte 3

Atributo	Média		Desvio padrão		Variância		Curtose		Assimetria		Percentil 95%	
	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes	Rotulados	Pendentes
APP_ESCADINHA_LAGO_NATURAL	0,03	0,01	1,03	0,26	1,06	0,07	10648,24	64892,96	91,34	183,92	0,00	0,00
APP_ESCADINHA_NASCENTE_OLHO_DAGUA	0,01	2,76E-03	0,06	0,04	3,77E-03	1,74E-03	1596,13	51627,44	34,61	166,34	0,00	0,00
APP_RIO_ATE_10	20,40	2,87	63,77	16,90	4066,11	285,58	638,85	5622,87	16,36	46,67	95,67	11,69
APP_RIO_10_A_50	0,84	0,20	27,87	3,28	776,84	10,78	22383,96	22987,35	145,05	92,53	0,00	0,00
APP_RIO_50_A_200	0,38	0,07	12,05	2,62	145,32	6,89	5417,61	154749,62	66,08	261,87	0,00	0,00
APP_RIO_200_A_600	0,34	0,02	21,93	1,97	480,87	3,90	16651,06	90831,75	120,64	247,92	0,00	0,00
APP_RIO_ACIMA_600	0,99	0,03	46,27	3,35	2141,36	11,22	5915,14	261256,38	68,87	400,58	0,00	0,00
APP_NASCENTE_OLHO_DAGUA	0,57	0,23	3,16	2,21	10,00	4,89	1043,84	13975,81	25,58	82,33	2,36	0,00
APP_LAGO_NATURAL	1,27	0,09	33,42	2,49	1117,01	6,21	7333,18	66965,09	81,38	184,79	4,92	0,00
APP_RESERVATORIO_ARTIFICIAL_DECORRENTE_BARRAMENTO	0,07	0,11	1,45	3,55	2,10	12,60	2400,60	441817,15	43,46	611,42	0,00	0,04
APP_VEREDA	0,04	0,07	2,35	5,64	5,54	31,83	11798,20	569445,02	100,32	596,50	0,00	0,00
APP_AREA_TOPO_MORRO	0,08	0,07	2,79	4,14	7,80	17,11	4267,12	128587,61	58,07	258,61	0,00	0,00
APP_MANGUEZAL	0,02	0,02	2,46	2,76	6,04	7,63	20008,02	223042,37	137,57	400,65	0,00	0,00
APP_AREA_ALTITUDE_SUPERIOR_1800	1,02E-05	1,90E-03	1,21E-03	0,71	1,47E-06	0,50	16513,16	564348,15	125,78	663,28	0,00	0,00
APP_BORDA_CHAPADA	0,35	0,09	12,70	8,38	161,28	70,21	5068,45	152195,44	64,39	309,84	0,00	0,00
APP_RESTINGA	0,00	0,01	0,45	7,38	0,20	54,49	23968,37	1842765,45	152,21	1299,21	0,00	0,00
APP_AREA_DECLIVIDADE_MAIOR_45	0,05	0,07	2,58	5,42	6,66	29,33	14476,37	284293,53	109,78	433,97	0,00	0,00
AREA_NAO_CLASSIFICADA	404,91	20,71	11142,14	489,89	124147391,30	239995,07	6417,51	544939,03	404,91	629,71	411,23	62,33
CORPO_DAGUA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
APP_AREA_VN	3,21	1,08	47,67	15,43	2272,17	238,21	4086,44	25985,80	54,36	114,60	2,94	2,90
APP_VAZIO	1,08	0,46	20,46	10,33	418,44	106,77	2929,04	58748,20	47,46	176,86	0,01	0,63
BANHADO	0,11	0,03	5,23	4,45	27,37	19,82	3551,27	519709,53	57,29	611,59	0,00	0,00
APP_BANHADO	0,11	0,03	5,23	4,44	27,35	19,75	3554,13	522044,86	57,31	612,86	0,00	0,00
AREA_IMOVEL_LIQUIDA_ANALISE	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ARL_AVERBADA_OUTRO_IMOVEL	0,04	0,05	2,61	8,07	6,80	65,16	12908,78	215354,63	106,02	402,22	0,00	0,00
SEDE_IMOVEL	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
RESERVATORIO_GERACAO_ENERGIA_ATE_24_08_2001	0,01	1,21E-03	0,82	0,28	0,67	0,08	15944,17	226093,64	180,27	419,96	0,00	0,00
APP_RESERVATORIO_GERACAO_ENERGIA_ATE_24_08_2001	0,01	1,10E-03	0,82	0,26	0,67	0,07	15944,17	255277,84	123,84	437,06	0,00	0,00
AREA_TERRITORIO_PCT	3,10	0,03	498,82	11,59	248818,38	134,29	25965,00	506230,26	161,14	639,47	0,00	0,00

Fonte: Do Autor (2022).

Com os resultados obtidos, foi constatada a presença de atributos constantes em zero, sendo estes: ‘AREA USO RESTRITO PANTANEIRA’; ‘NASCENTE OLHO DAGUA’; ‘ARL A RECUPERAR’; ‘APP A RECUPERAR’; ‘CORPO DAGUA’; ‘AREA IMOVEL LIQUIDA ANALISE’ e ‘SEDE IMOVEL’. Para a base rotulada, foram encontrados 70 atributos com elevada assimetria e, para a base de não rotulada (pendentes), o número de *features* com elevada assimetria foi de 71. Outro ponto a ser levantado, é a possibilidade de *overflow*[1] para algumas medidas, devido ao elevado valor encontrado para algumas medidas (valores na ordem de 10^{90}, por exemplo), o que desponta como valor discrepante, analisando, conjuntamente, com as medidas de percentil.

Comparando as medidas estatísticas descritivas entre os registros rotulados e não rotulados, observou-se uma diferença de valores significativa entre as bases de dados. Assim, para esta primeira análise, não seria possível o uso dos atributos pendentes para uma abordagem semi-supervisionada. O uso de estudos comparativos mais profundos entre os conjuntos é recomendada, haja vista que o número de cadastros sem rotulação é 71 vezes maior que o número de registros já analisados e rotulados. Portanto, de acordo com as observações feitas por meio das estatísticas descritivas das bases de dados, não foi possível encontrar uma viabilidade para uso dos registros pendentes. Recomenda-se análises mais profundas, em que o autor dessa Dissertação pretende realizar em trabalhos futuros.

Além do levantamento dos valores das estatísticas descritivas, foram gerados, para 10 atributos, gráficos de dispersão entre pares (*pairplots*) e histogramas. Os atributos escolhidos foram os que apresentaram uma maior relevância durante os estudos ao longo desta dissertação, sendo estes: ‘AREA DOC’; ‘QTD RETIFICACOES’; ‘APP RESERVATORIO ARTIFICIAL DECORRENTE BARRAMENTO’; ‘APP LAGO NATURAL’; ‘APP ESCADINHA’; ‘APP ESCADINHA NASCENTE OLHO DAGUA’; ‘QTD SOB IR’; ‘DESEJA ADERIR PRA’; ‘EXISTE TAC’ e ‘AREA CONSOLIDADA’. Para uma melhor visualização, os *pairplots* foram divididos em 2 grupos, inseridos nas Figuras 4.1 e 4.2. Além disto, para o eixo horizontal do gráfico, os nomes dos atributos foram codificados de ‘X_1’ à ‘X_10’ para fins de ajuste da legenda no gráfico. Na Tabela **??** é apresentada a legenda para os atributos codificados.

Para uma melhor visualização das distribuições de cada atributo, os histogramas foram gerados à parte, de maneira a fornecer um maior destaque para as análises visuais. Os histogramas foram gerados para os mesmos atributos utilizados para obtenção dos *pairplots* e seguem dispostos nas Figuras 4.3, 4.4 e 4.5.

[1] termo aplicado para a ocorrência do valor de um dado número ultrapassar o valor limite de operação da máquina, ou seja, quando o expoente da variável é maior que o expoente suportado pelo sistema.

Analisando os *pairplots* conjuntamente com os histogramas, alguns pontos cabem destaque: os atributos numéricos possuem grande assimetria em sua distribuição, o que pode ser visto nos atributos 'AREA DOC'; 'QTD RETIFICACOES'; 'APP RESERVATORIO ARTIFICIAL DECORRENTE BARRAMENTO' e 'APP LAGO NATURAL'; 'APP ESCADINHA'; 'APP ESCADINHA NASCENTE OLHO DAGUA' e 'QTD SOB IR' e 'AREA CONSOLIDADA'. Nos quais, a maioria dos atributos apresentam maior concentração nos baixos valores. Tal fato também indica a presença de *outliers*, o que pode ser visto nos gráficos de dispersão. Para atributos de área, a presença destes valores elevados indica a presença de alguma grande propriedade, o que, apesar de deslocado dos demais valores, não precisamente pode indicar um valor incorreto. Por fim, cabe salientar que os gráficos de dispersão apontam que as classes não são separáveis. A não separabilidade entre as classes pode ser observada tanto pelos gráficos de dispersão, quanto pelos histogramas, onde não é possível discriminar uma classe da outra de forma visual.

Tabela 4.4 – Codificação dos atributos para os gráficos de *pairplot*

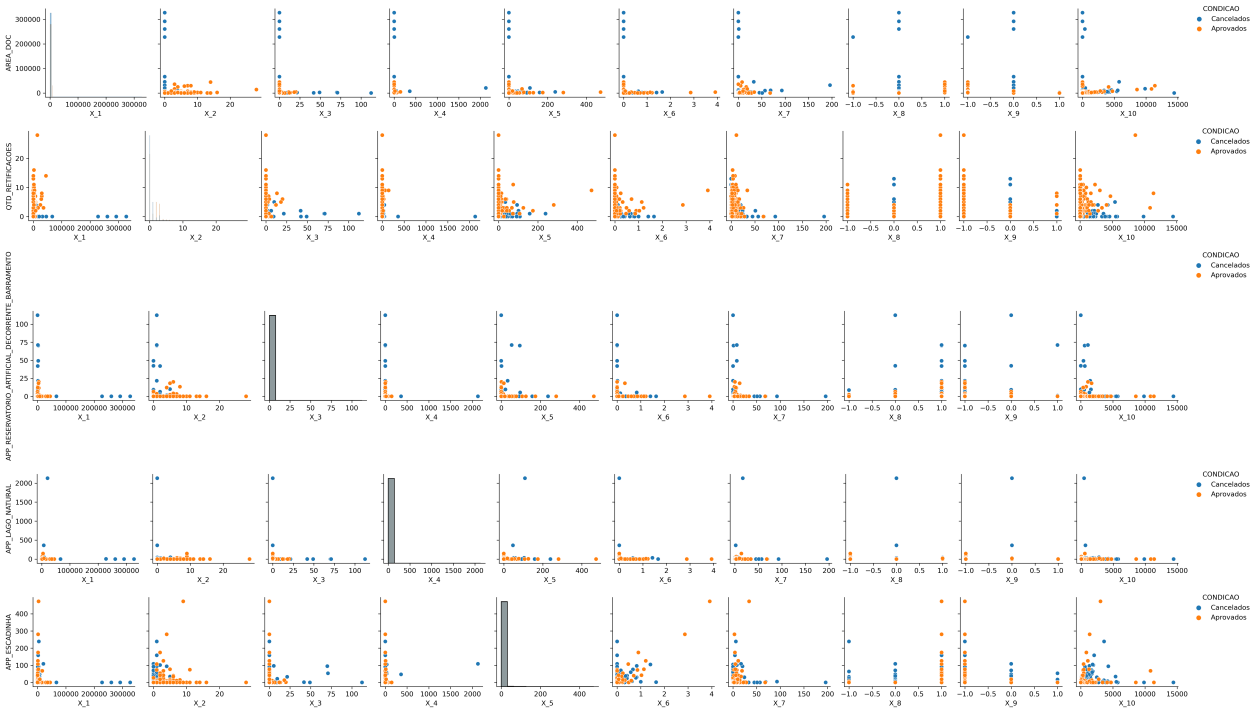
Atributo	Nome Codificado
AREA DOC	X_1
QTD RETIFICACOES	X_2
APP RESERVATORIO ARTIFICIAL DECORRENTE BARRAMENTO	X_3
APP LAGO NATURAL	X_4
APP ESCADINHA	X_5
APP ESCADINHA NASCENTE OLHO DAGUA	X_6
QTD SOB IR	X_7
DESEJA ADERIR PRA	X_8
EXISTE TAC	X_9
AREA CONSOLIDADA	X_10

Fonte: Do Autor (2022).

4.2 Resultados dos ensaios comparativos para a seleção de atributos

Após os procedimentos de limpeza e codificação da base de dados, remoção de atributos constantes e atributos sensíveis e/ou irrelevantes para a classificação, foram realizados os procedimentos de seleção de *features* por meio do FDR. Os testes realizados com o FDR, variando o número de atributos, apontaram 49 *features* como o melhor resultado preditivo. Juntamente com a seleção de atributos via FDR, foi aplicada uma remoção de atributos altamente correlacionados, por meio da matriz de correlação de Pearson. Desta forma, para os pares de atributos com correlação maior que 0,9, um dos atributos foi eliminado, tal procedimento foi realizado para fins de verificação do impacto de atributos correlacionados na classificação. Assim, foram

Figura 4.1 – Gráficos de dispersão e histogramas para os atributos ‘AREA DOC’; ‘QTD RETIFICACOES’; ‘APP RESERVATORIO ARTIFICIAL DECORRENTE BARRAMENTO’; ‘APP LAGO NATURAL’ e ‘APP ESCADINHA’.

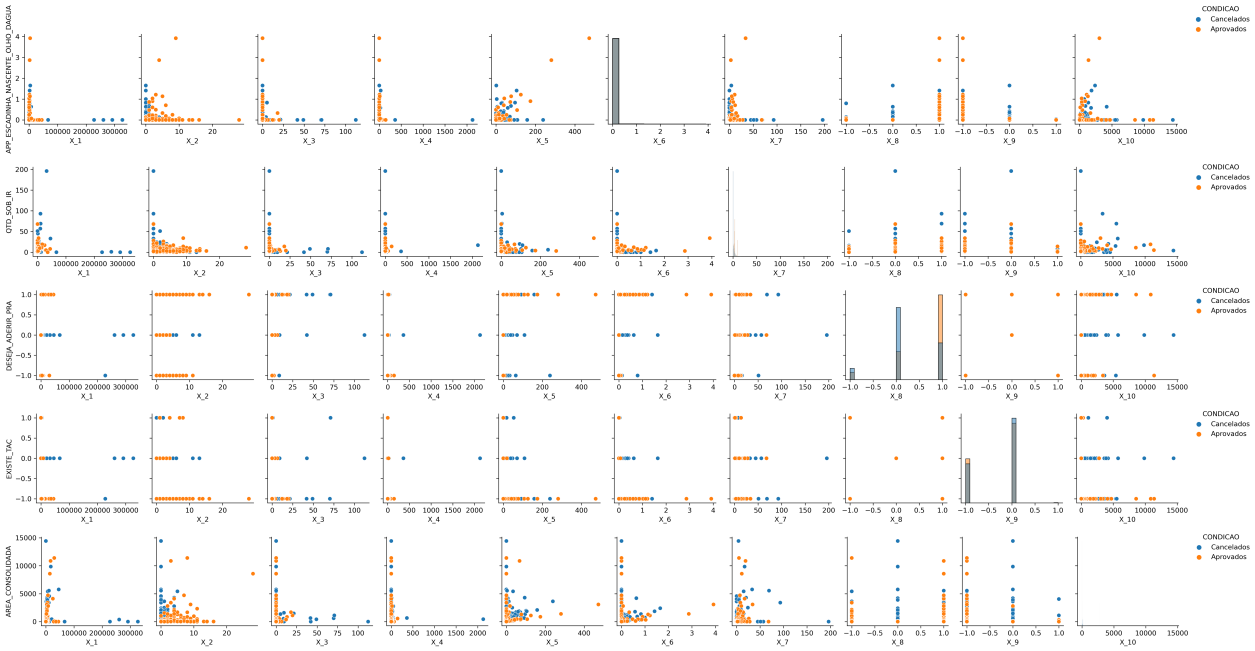


Fonte: Fonte: Do Autor (2022).

eliminados 8 atributos, gerando uma segunda base de dados com seleção de *features* com 41 variáveis de entrada. A Figura 4.6 mostra a matriz de correlação contendo as 49 variáveis selecionadas primeiramente. Os 8 atributos correlacionados que foram removidos por alta correlação possuem correlação prática com atributos que foram mantidos na base de dados.

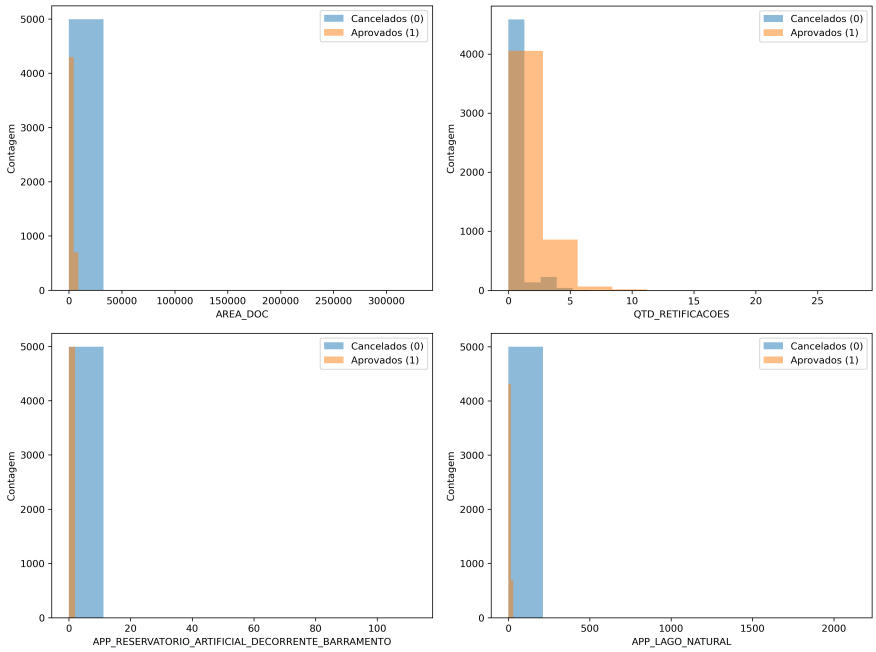
Os atributos ‘AREA HA’, ‘AREA IMOVEL LIQUIDA’, ‘AREA NAO CLASSIFICADA’ e ‘AREA IMOVEL’ são correlacionados com o atributo ‘NUMERO MF’, o que possui correspondência prática, uma vez que ‘NUMERO MF’ corresponde ao número de módulos fiscais, uma medida de área específica de cada município e os demais atributos se referem à área total do imóvel rural. Os atributos ‘APP RESTINGA’ e ‘RESTINGA’ são correlacionados, uma vez que a área de proteção permanente (APP) de uma restinga (‘APP RESTINGA’) é proporcional a área de restinga de um imóvel rural. A mesma analogia pode ser feito com o par correlacionado ‘APP AREA TOPO MORRO’ e ‘AREA TOPO MORRO’. Os atributos ‘EXISTE TAC’ e ‘EXISTE PRAD’ possuem correlação prática uma vez que tanto o TAC (Termo de Ajustamento de Conduta) quando o PRAD (Projeto de Recuperação de Áreas Degradadas) são documentos que auxiliam na

Figura 4.2 – Gráficos de dispersão e histogramas para os atributos ‘APP ESCADINHA NASCENTE OLHO DAGUA’; ‘QTD SOB IR’; ‘DESEJA ADERIR PRA’; ‘EXISTE TAC’ e ‘AREA CONSOLIDADA’.



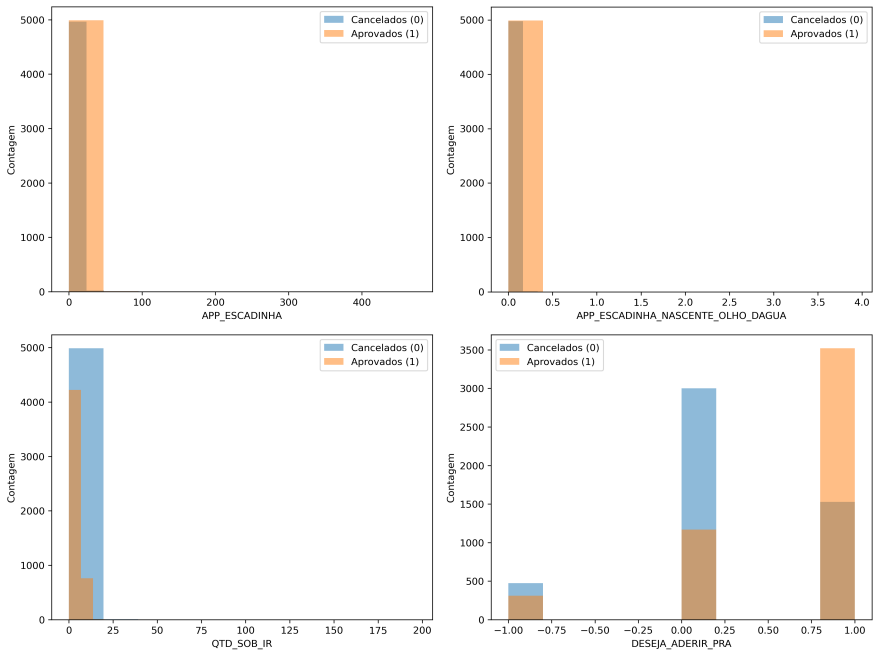
Fonte: Fonte: Do Autor (2022).

Figura 4.3 – Histogramas para os atributos ‘AREA DOC’; ‘QTD RETIFICACOES’; ‘APP RESERVATORIO ARTIFICIAL DECORRENTE BARRAMENTO’ e ‘APP LAGO NATURAL’.



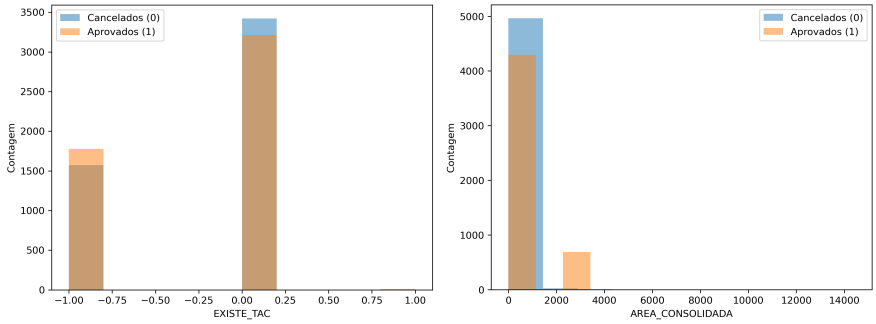
Fonte: Fonte: Do Autor (2022).

Figura 4.4 – Histogramas para os atributos ‘APP ESCADINHA’; ‘APP ESCADINHA NASCENTE OLHO DAGUA’; ‘QTD SOB IR’ e ‘DESEJA ADERIR PRA’.



Fonte: Fonte: Do Autor (2022).

Figura 4.5 – Histogramas para os atributos ‘EXISTE TAC’ e ‘AREA CONSOLIDADA’.

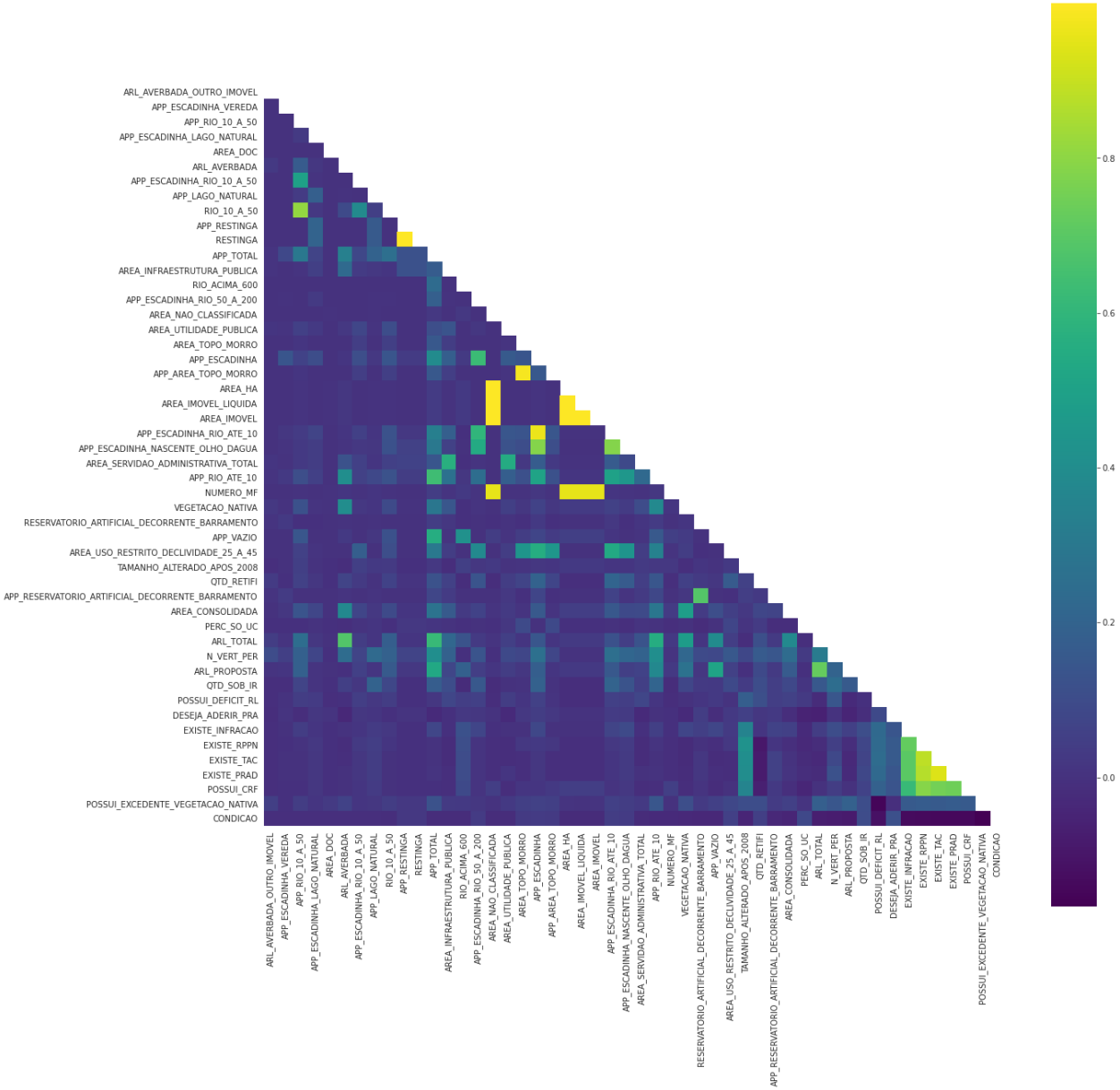


Fonte: Fonte: Do Autor (2022).

regularização ambiental da propriedade. Os atributos ‘APP ESCADINHA RIO ATE 10’ e ‘APP ESCADINHA’ correspondem a valores de área de proteção permanente que deve ser preservada dentro da propriedade rural.

Realizados os procedimentos de seleção de atributos, foram gerados 3 subconjuntos de dados dentro da base de dados original, sendo estes: sem seleção de atributos, com seleção realizada pelo FDR com e sem a remoção de atributos via correlação. A Tabela 4.5 apresenta o número de atributos para cada subconjunto de dados.

Figura 4.6 – Matriz de Correlação para o conjunto de dados de treinamento.



Fonte: Fonte: Do Autor (2021).

Tabela 4.5 – Número de atributos utilizados em cada método de seleção de *features* aplicado

Método de seleção	Sem seleção	FDR	FDR + Correlação
Número de atributos	83	49	41

Fonte: Do Autor (2022).

Após a aplicação dos métodos de seleção de atributos, foi realizado o procedimento de treinamento e teste sob validação cruzada com os classificadores. A variação dos hiperparâmetros foi realizada respeitando a faixa de valores estipulada na Tabela 3.4. Os valores de cada hiperparâmetro para cada modelo em cada método de seleção de atributos estão dispostos na Tabela 4.6. Nas Redes MLP, foi-se utilizada a função *softmax* na camada de saída, de forma a permitir, além da classificação por rótulo (0 ou 1), a estimação da probabilidade a *posteriori* de cada registro pertencer à cada uma das classes. Tal medida é importante para o cálculo da curva ROC e a obtenção da medida de AUC.

Tabela 4.6 – Hiperparâmetros utilizados pelos classificadores para cada tipo de subconjunto de dados

Modelo	Hiperparâmetro	Sem Seleção	FDR	FDR + Correlação
ABC	Medida de avaliação	gini	gini	gini
	Divisão de folhas	'best'	'best'	'best'
	Profundidade máxima	7	11	9
	Número de árvores	400	100	300
GBT	Profundidade máxima	9	9	5
	Taxa de aprendizado	0,1	0,05	0,2
	Número de árvores	300	250	300
LRC	C	29,7635	78,47	1438,45
	Solver	newton-cg	newton-cg	lbfgs
	Regularização	nenhuma	nenhuma	l2
MLP	Neurônios na camada oculta	15	(10, 15)	25
	ativacao	tangente hiperbólica	relu	tangente hiperbólica
	Solver	lbfgs	lbfgs	lbfgs
	Taxa de aprendizado	constante	adaptativo	adaptativo
RFC	Profundidade máxima	19	17	19
	Medida de avaliação	entropia	entropia	entropia
	Número de árvores	300	400	100
SVM	C	500	500	1000
	γ	auto	auto	scale

Fonte: Do Autor (2022).

Obtidos os hiperparâmetros de melhor desempenho, os classificadores foram treinados e testados sob validação cruzada e tendo sua acurácia comparada pelo teste t de *Student*. Por fim, os classificadores foram aplicados ao conjunto de validação. Os resultados preditivos estão dispostos na Tabela 4.7. Vale ressaltar que o conjunto de teste se refere aos resultados preditivos relativos aos conjuntos de teste gerados pela validação cruzada e estão dispostos no formato média $\pm$ desvio-padrão, enquanto o conjunto de validação se refere ao conjunto de registros novos aos classificadores.

Tabela 4.7 – Resultados de desempenho dos modelos de aprendizagem relativos à cada método de seleção de atributos

Método de seleção	Modelo	ACC	AUC	*F1-Score*	*Precision* - 0	*Precision* - 1	*Recall* - 0	*Recall* - 1
Teste Sem seleção	ABC	96,4 ± 0,4	98,7 ± 0,3	97,5 ± 0,2	97,3 ± 0,3	90,8 ± 1,2	97,8 ± 0,3	88,7 ± 1,4
	GBT	**96,5 ± 0,4**	99,2 ± 0,2	**97,8 ± 0,2**	97,3 ± 0,3	**92,7 ± 1,1**	**98,3 ± 0,3**	89,0 ± 1,7
	LRC	90,5 ± 0,4	92,7 ± 0,5	94,3 ± 0,3	92,2 ± 0,3	81,9 ± 1,8	96,5 ± 0,4	66,0 ± 1,6
	MLP	93,5 ± 0,4	97,4 ± 0,3	96,0 ± 0,2	94,5 ± 0,4	88,5 ± 1,4	97,6 ± 0,3	76,4 ± 1,9
	RFC	**96,5 ± 0,4**	**99,3 ± 0,1**	**97,8 ± 0,2**	**98,1 ± 0,3**	90,2 ± 1,4	97,6 ± 0,4	92,2 ± 1,4
	SVM	88,7 ± 0,7	95,2 ± 0,5	92,6 ± 0,4	98,0 ± 0,3	64,7 ± 1,4	87,8 ± 0,8	**92,4 ± 1,1**
Validação Sem seleção	ABC	96,6	98,9	97,9	97,7	92,0	98,1	90,6
	GBT	**97,1**	**99,3**	**98,2**	97,7	**94,2**	**98,7**	90,6
	LRC	90,5	92,7	94,2	92,1	81,8	96,5	65,9
	MLP	94,0	97,7	96,3	94,8	89,8	97,8	77,9
	RFC	**97,1**	**99,3**	**98,2**	**98,5**	91,3	97,8	**94,0**
	SVM	89,5	95,3	93,2	98,2	66,5	88,7	93,3
Teste FDR	ABC	95,9 ± 0,4	98,9 ± 0,2	97,4 ± 0,2	97,1 ± 0,3	90,7 ± 1,5	97,8 ± 0,4	88,1 ± 1,1
	GBT	96,4 ± 0,4	**99,3 ± 0,1**	**97,8 ± 0,2**	97,4 ± 0,4	**92,1 ± 1,3**	**98,2 ± 0,3**	89,2 ± 1,7
	LRC	87,3 ± 0,7	90,5 ± 1,0	92,4 ± 0,4	89,3 ± 0,5	74,9 ± 2,5	95,8 ± 0,5	52,3 ± 2,6
	MLP	93,3 ± 0,4	97,4 ± 0,4	95,9 ± 0,2	94,1 ± 0,6	89,6 ± 1,3	97,9 ± 0,3	74,5 ± 2,8
	RFC	**96,5 ± 0,3**	**99,3 ± 0,1**	**97,8 ± 0,2**	**98,2 ± 0,2**	89,8 ± 1,0	97,4 ± 0,3	**92,6 ± 1,0**
	SVM	88,8 ± 0,6	95,4 ± 0,6	92,7 ± 0,4	98,0 ± 0,3	65,0 ± 1,4	87,9 ± 0,8	**92,6 ± 1,0**
Validação FDR	ABC	96,7	98,9	97,9	97,7	92,4	98,2	90,3
	GBT	**97,0**	**99,3**	**98,1**	97,8	**93,5**	**98,5**	90,7
	LRC	86,9	90,0	92,1	89,2	72,9	95,3	52,3
	MLP	93,7	97,4	96,1	94,4	89,8	97,9	76,2
	RFC	**97,0**	**99,3**	**98,1**	**98,5**	91,0	97,8	**94,0**
	SVM	89,7	95,8	93,3	98,2	66,9	88,8	93,4
Teste FDR + Correlação	ABC	95,8 ± 0,4	98,7 ± 0,3	97,4 ± 0,2	97,1 ± 0,4	90,1 ± 1,1	97,8 ± 0,3	87,9 ± 1,9
	GBT	**96,3 ± 0,5**	99,1 ± 0,2	**97,7 ± 0,3**	97,2 ± 0,4	**92,4 ± 1,8**	**98,2 ± 0,4**	88,2 ± 1,6
	LRC	85,0 ± 0,7	89,9 ± 0,6	91,2 ± 0,4	86,6 ± 0,8	71,1 ± 2,0	96,3 ± 0,3	38,2 ± 4,5
	MLP	93,2 ± 0,3	97,1 ± 0,3	95,8 ± 0,2	94,4 ± 0,4	87,3 ± 1,3	97,3 ± 0,4	76,2 ± 2,1
	RFC	96,2 ± 0,4	**99,2 ± 0,1**	**97,7 ± 0,3**	**98,0 ± 0,4**	89,4 ± 1,5	97,4 ± 0,4	91,7 ± 1,8
	SVM	88,7 ± 0,5	95,0 ± 0,4	92,6 ± 0,3	**98,0 ± 0,2**	64,7 ± 1,2	87,8 ± 0,7	**92,5 ± 0,9**
Validação FDR+ Correlação	ABC	96,6	98,8	97,9	97,7	92,2	98,2	90,4
	GBT	**96,9**	99,2	**98,1**	97,6	93,5	**98,5**	90,2
	LRC	84,7	89,3	91,0	86,3	70,2	96,2	36,8
	MLP	93,6	97,3	96,1	94,8	88,2	97,5	77,7
	RFC	96,7	**99,3**	97,9	**98,3**	90,1	97,5	93,2
	SVM	89,4	95,2	93,1	98,2	66,2	88,5	**93,3**

Fonte: Do Autor (2022).

A partir dos resultados obtidos nos ensaios de classificação, comparando os resultados preditivos em relação aos métodos de seleção de atributos, podem ser levantados alguns pontos importantes. Primeiramente, para as medidas de ACC, AUC e *F1-Score*, os classificadores ABC, GBT e RFC atingiram valores acima de 95% para o conjunto de validação. Para as medidas de *Precision* e *Recall*, os valores para os mesmos 3 classificadores para o conjunto de validação ficou acima de 90%. Os classificadores LRC, MLP e SVM atingiram valores acima de 89% para as medidas de AUC, ACC e *F1-Score*, entretanto, para as medidas de *Precision* e *Recall* é observado um desempenho inferior quando comparados com o ABC, GBT e RFC. Outro ponto a ser destacado é a sensibilidade dos classificadores em relação ao subconjunto de dados utilizado. O

classificador LRC apresentou maior flutuação no desempenho preditivo conforme foram aplicadas mudanças nos atributos utilizados, enquanto os demais modelos obtiveram diferenças menos significativas, sobretudo quando analisados os resultados para o conjunto de validação.

Os testes *t* de *Student* foram aplicados sobre os classificadores com o objetivo de analisar se a acurácia gerada por um classificador possui ou não diferença estatística significativa. Os resultados do teste t de *Student* para os conjuntos de dados sem seleção de atributos, com seleção de atributos via FDR, com seleção de atributos e via FDR + correlação são apresentados nos Quadros 4.1, 4.2 e 4.3, respectivamente. Os valores destacados se referem aos valores em que a hipótese nula é rejeitada, ou seja, aos testes que apresentam diferença estatística significativa entre as acurácias dos classificadores, dada por $p < 0,05$. Analisando os resultados dos testes estatísticos para os modelos treinados sem seleção de atributos, os pares de classificadores ABC x GBT e RFC x GBT não obtiveram resultados com diferença estatística significativa, tendo os modelos destes pares obtidos os melhores resultados de classificação. Em relação aos modelos treinados com seleção de atributos por meio do FDR, os pares de classificadores ABC x GBT, RFC x ABC, RFC x GBT e SVM x LRC não obtiveram diferença significativa, sendo os classificadores ABC, RFC e GBT aqueles que obtiveram os melhores resultados. Sobre os classificadores treinados com a seleção de atributos via FDR + Correlação, os pares de classificadores que não apresentaram diferença significativa pelo teste *t* foram RFC x ABC, RFC x GBT e SVM x LRC, sendo os dois primeiros pares (RFC x ABC e RFC x GBT) com os melhores resultados (classificadores ABC, RFC e GBT).

Quadro 4.1 – Resultados do teste *t* de *Student* para cada par de classificadores treinados sem seleção de atributos

	ABC	GBT	LRC	MLP	RFC
GBT	0,144				
LRC	0,000	0,000			
MLP	0,000	0,000	0,000		
RFC	0,011	0,609	0,000	0,000	
SVM	0,000	0,000	0,000	0,000	0,000

Fonte: Do Autor (2022).

Além destes resultados, foram geradas as curvas ROC para o conjunto de validação. As Figuras 4.7, 4.8 e 4.9 apresentam as curvas ROC para os classificadores ABC, GBT, LRC, MLP e SVM para os dados sem seleção de atributos, com seleção de atributos via FDR e com seleção de atributos via FDR + Correlação, respectivamente. A visualização das Curvas ROC corrobora com os resultados numéricos gerados, tanto pela Tabela 4.7 quanto pelos Quadros 4.1, 4.2 e 4.3, onde as Curvas ROC do ABC, GBT e RFC se

Quadro 4.2 – Resultados do teste *t* de *Student* para cada par de classificadores treinados com seleção de atributos via FDR

	ABC	GBT	LRC	MLP	RFC
GBT	0,106				
LRC	0,004	0,004			
MLP	0,000	0,000	0,028		
RFC	0,066	0,603	0,003	0,000	
SVM	0,000	0,000	0,509	0,000	0,000

Fonte: Do Autor (2022).

Quadro 4.3 – Resultados do teste *t* de *Student* para cada par de classificadores treinados com seleção de atributos via FDR + Correlação

	ABC	GBT	LRC	MLP	RFC
GBT	0,020				
LRC	0,003	0,003			
MLP	0,000	0,000	0,012		
RFC	0,065	0,701	0,002	0,000	
SVM	0,000	0,000	0,114	0,001	0,000

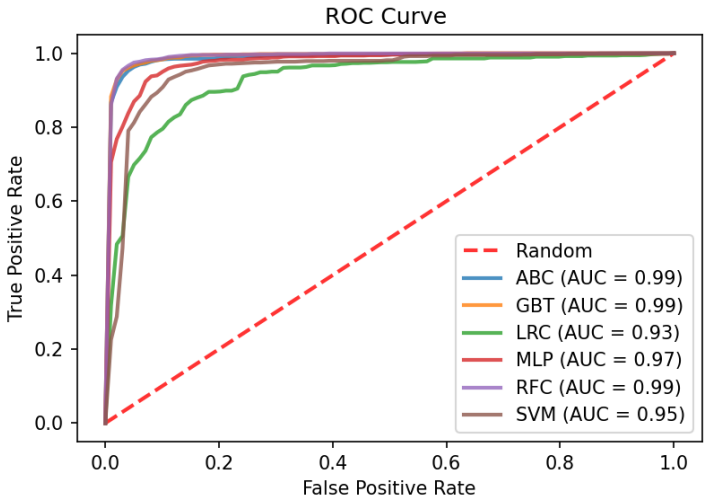
Fonte: Do Autor (2022).

sobrepõem em valores acima dos modelos LRC, MLP e SVM, mostrando um desempenho superior. Este efeito pode ser observado ao longo dos 3 métodos de seleção de atributos. Além da sobreposição, a ordem do desempenho dos classificadores, por meio da análise da Curva ROC se mantém a mesma, sendo a seguinte: o LRC apresentou o menor desempenho, seguido pelo SVM e a MLP com ligeira superioridade, enquanto os classificadores ABC, GBT e RFC se mostraram equivalentes.

De maneira a simplificar a visualização dos efeitos da seleção de atributos sobre cada um dos 3 classificadores de melhor desempenho (ABC, GBT e RFC), foi gerada a Tabela 4.8 para verificar se há diferença numérica que possa indicar algum método de seleção de atributos que se sobressaia no desempenho preditivo. Realizando um comparativo entre os métodos de seleção de atributos em cada classificador e associando com os resultados de número de *features* obtidas para cada método de seleção vistos na Tabela 4.5, pode-se apontar um método de seleção de atributos mais eficiente para a aplicação. Para isto, são analisados, conjuntamente, os resultados de desempenho desempenho preditivo dos classificadores, por meio das medidas de avaliação e no número de atributos apontado em cada método de seleção. Desta forma, o melhor subconjunto foi determinado como aquele de menor número de *features* e melhor desempenho.

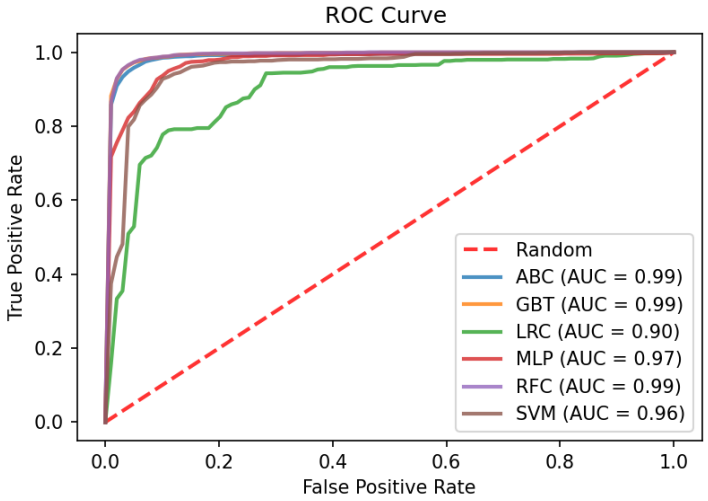
Observando os resultados da Tabela 4.8, verifica-se que uma baixa diferença percentual entre um mesmo classificador sob diferentes métodos de seleção de atributos, diferença esta, na maioria dos casos,

Figura 4.7 – Curva ROC para os classificadores treinados sem seleção de atributos.



Fonte: Fonte: Do Autor (2022).

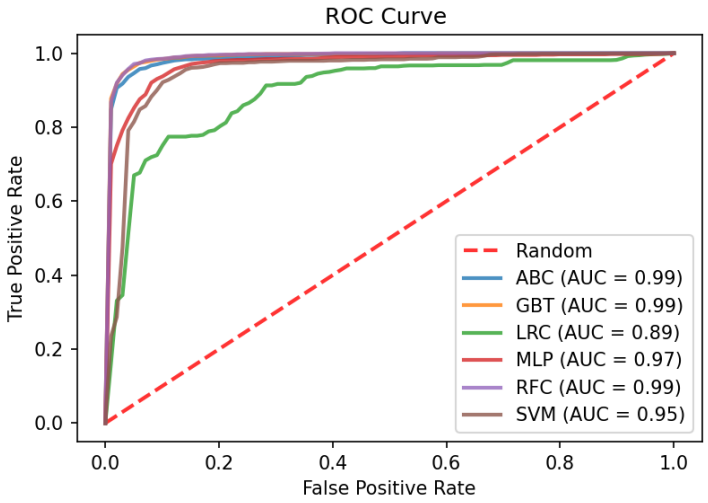
Figura 4.8 – Curva ROC para os classificadores treinados com seleção de atributos via FDR.



Fonte: Fonte: Do Autor (2022).

menor que 1% quando comparados o menor e o maior valor da medida de desempenho, sobretudo, quando dado enfoque aos resultados para o conjunto de validação. Assim, baseando-se somente no critério de desempenho preditivo, não é possível apontar qual método se sobressai, assim, sendo necessário, avaliar o número de atributos utilizados por cada método de seleção para se escolher um método específico, buscando-se modelos mais parcimoniosos. Conforme o menor número de atributos selecionados foi obtido pelo método de

Figura 4.9 – Curva ROC para os classificadores treinados com seleção de atributos via FDR + Correlação.



Fonte: Fonte: Do Autor (2022).

Tabela 4.8 – Comparativo dos métodos de seleção de atributos em cada um dos 3 melhores classificadores

Conjunto	Modelo	ACC	AUC	*F1-Score*	*Precision* - 0	*Precision* - 1	*Recall* - 0	*Recall* - 1
Teste	ABC (Sem Seleção)	**96,4 ± 0,4**	98,7 ± 0,3	**97,5 ± 0,2**	**97,3 ± 0,3**	**90,8 ± 1,2**	**97,8 ± 0,3**	**88,7 ± 1,4**
	ABC (FDR)	95,9 ± 0,4	**98,9 ± 0,2**	97,4 ± 0,2	97,1 ± 0,3	90,7 ± 1,5	**97,8 ± 0,4**	88,1 ± 1,1
	ABC (FDR + Correlação)	95,8 ± 0,4	98,7 ± 0,3	97,4 ± 0,2	97,1 ± 0,4	90,1 ± 1,1	**97,8 ± 0,3**	87,9 ± 1,9
	GBT (Sem Seleção)	**96,5 ± 0,4**	99,2 ± 0,2	**97,8 ± 0,2**	97,3 ± 0,3	**92,7 ± 1,1**	**98,3 ± 0,3**	89,0 ± 1,7
	GBT (FDR)	96,4 ± 0,4	**99,3 ± 0,1**	**97,8 ± 0,2**	**97,4 ± 0,4**	92,1 ± 1,3	98,2 ± 0,3	**89,2 ± 1,7**
	GBT (FDR + Correlação)	96,3 ± 0,5	99,1 ± 0,2	97,7 ± 0,3	97,2 ± 0,4	92,4 ± 1,8	98,2 ± 0,4	88,2 ± 1,6
	RFC (Sem Seleção)	**96,5 ± 0,4**	**99,3 ± 0,1**	**97,8 ± 0,2**	98,1 ± 0,3	**90,2 ± 1,4**	**97,6 ± 0,4**	92,2 ± 1,4
	RFC (FDR)	**96,5 ± 0,3**	**99,3 ± 0,1**	**97,8 ± 0,2**	**98,2 ± 0,2**	89,8 ± 1,0	97,4 ± 0,3	**92,6 ± 1,0**
	RFC (FDR + Correlação)	96,2 ± 0,4	99,2 ± 0,1	97,7 ± 0,3	98,0 ± 0,4	89,4 ± 1,5	97,4 ± 0,4	91,7 ± 1,8
Validação	ABC (Sem Seleção)	96,6	**98,9**	**97,9**	**97,7**	92,0	98,1	**90,6**
	ABC (FDR)	**96,7**	98,9	**97,9**	**97,7**	**92,4**	**98,2**	90,3
	ABC (FDR + Correlação)	96,6	98,8	**97,9**	**97,7**	92,2	**98,2**	90,4
	GBT (Sem Seleção)	**97,1**	**99,3**	**98,2**	97,7	**94,2**	**98,7**	90,6
	GBT (FDR)	97,0	**99,3**	98,1	**97,8**	93,5	98,5	**90,7**
	GBT (FDR + Correlação)	96,9	99,2	98,1	97,6	93,5	98,5	90,2
	RFC (Sem Seleção)	**97,1**	**99,3**	**98,2**	**98,5**	**91,3**	**97,8**	**94,0**
	RFC (FDR)	97,0	**99,3**	98,1	**98,5**	91,0	**97,8**	**94,0**
	RFC (FDR + Correlação)	96,7	**99,3**	97,9	98,3	90,1	97,5	93,2

Fonte: Do Autor (2022).

FDR + Correlação, pode-se apontá-lo como o mais eficiente pelo menor número de atributos utilizados e ser escolhido o seu subconjunto como o subconjunto a ser utilizado para os testes posteriores. Contudo, para o caso de considerar somente o desempenho preditivo, ambos os métodos de seleção poderão ser utilizados, pois a seleção de atributos não impactou significativamente no desempenho preditivo para esta aplicação.

4.3 Resultados dos ensaios comparativos para os métodos de *oversampling*

Após a escolha do conjunto de atributos a ser utilizado, foram realizados testes aplicando *oversampling* sobre a base de dados a fim de mitigar o desbalanceamento entre as classes. Tal desbalanceamento pode ser observado na Tabela 3.3, onde a classe de Cancelados é, aproximadamente, 4 vezes mais numerosa que a classe de Aprovados. Visando balancear o conjunto de treinamento, foram aplicados dois procedimentos de superamostragem: a Reamostragem Aleatória e o SMOTE. Desta forma, o resultado gerado na seção anterior utilizando o método de seleção FDR + Correlação será utilizado como *baseline*. Assim, será analisado se o *oversampling* aplicado sobre a base de dados refletiu em melhorias no desempenho preditivo. Conforme mencionado na seção 3.3, a reamostragem aleatória não possui hiperparâmetros de ajuste, este foi aplicado diretamente sobre a base de dados e utilizado para os ensaios de classificação. O SMOTE, após a variação do hiperparâmetro k, o valor $k = 9$ foi o que gerou os melhores resultados de acurácia. Os hiperparâmetros para os classificadores foram ajustados conforme as Tabelas 3.4 e 3.5 com os hiperparâmetros utilizados para o método de seleção de atributos tomado como *baseline*. Os hiperparâmetros ajustados para cada classificador em relação a cada método de superamostragem estão contidos na Tabela 4.9.

Tabela 4.9 – Hiperparâmetros utilizados nos ensaios de classificação que obtiveram maior acurácia durante o processo de ajuste

Modelo	Hiperparâmetro	*Baseline*	Reamostragem Aleatória	SMOTE
ABC	Medida de avaliação	gini	entropia	entropia
	Divisão de folhas	'best'	'best'	'best'
	Profundidade máxima	9	15	13
	Número de árvores	300	250	300
GBT	Profundidade máxima	5	15	7
	Taxa de aprendizado	0,2	0,15	0,2
	Número de árvores	300	500	300
LRC	C	1438,45	3792,6901	4,2813
	Solver	lbfgs	'newton-cg'	'newton-cg'
	Regularização	l2	l2	nenhuma
MLP	Neurônios na camada oculta	25	45	30
	ativacao	tangente hiperbólica	tangente hiperbólica	tangente hiperbólica
	solver	lbfgs	lbfgs	lbfgs
	Taxa de aprendizado	adaptativo	adaptativo	adaptativo
RFC	Profundidade máxima	19	19	17
	Medida de avaliação	entropia	entropia	gini
	Número de árvores	100	200	250
SVM	C	1000	1000	1000
	γ	'scale'	'scale'	'scale'

Fonte: Do Autor (2022).

Os resultados preditivos para os classificadores treinados sem superamostragem (*baseline*) e com os métodos de *oversampling* aplicados seguem na Tabela 4.10. Observando os resultados para o conjunto de teste, observa-se um aumento nos índices de desempenho para todos os classificadores, dando ênfase aos resultados obtidos pela Reamostragem Aleatória. Entretanto, tal ganho de desempenho não possui grande reflexo sobre o conjunto de validação, onde a diferença entre os índices gerados pelos classificadores não possuem diferença significativa.

Analisando as medidas de desempenho individualmente e dando enfoque ao melhor resultado obtido, as medidas globais (ACC, AUC e *F1-Score*) não possuem diferença significativa (diferenças menores que 0.5%), o que mostrou um baixo impacto dos métodos de superamostragem nestas medidas. Em relação às medidas relativas à cada classe, a Precisão para a classe 0 (Cancelados), a maior diferença entre o melhor desempenho do *baseline* e dos classificadores treinados com superamostragem foi de 0.5%, enquanto a precisão para a classe 1 (Aprovados) apresentou uma queda, para o *oversampling* realizado pelo SMOTE. Para as medidas de *Recall*, a classificação realizada em *baseline* apresentou melhor resultado na classe 0, enquanto para a classe 1, os classificadores treinados utilizando a superamostragem apresentaram ganhos de 1.2% e 1.9% para a Reamostragem Aleatória e para o SMOTE, respectivamente. Contudo, para uma contextualização geral dos resultados obtidos, o impacto do *oversampling* sobre a classificação não foi significativo de maneira que se possa justificar um aumento na base de dados com a inclusão de registros sintéticos para esta aplicação em específico.

Os resultados dos testes estatísticos para os classificadores treinados utilizado a Reamostragem Aleatória e o SMOTE estão dispostos, respectivamente nos Quadros 4.4 e 4.5. Analisando, primeiramente, os testes estatísticos realizados pelos classificadores treinados sob a Reamostragem aleatória, observa-se que os pares que não possuem diferença estatística significativa são os pares ABC x GBT e RFC x GBT, nos quais, os 3 modelos envolvidos possuem os melhores resultados preditivos de acordo com os números gerados pela Tabela 4.10, logo, pode-se inferir que estes modelos possuem o melhor desempenho preditivo sem que haja um modelo específico que se sobressaia estatisticamente. Analisando os resultados do teste t de *Student* para a classificação utilizando a superamostragem via SMOTE, somente o par RFC x GBT não possui diferença estatística significativa. Logo, como os dois classificadores apresentam os melhores resultados preditivos, para o treinamento via SMOTE pode-se inferir que os modelos GBT e RFC possuem os melhores resultados, contudo, analisando somente os resultados preditivos, o classificador ABC possui uma diferença, para algumas medidas, menor que 0.5%, o que pode manter o ABC como um dos modelos de melhor desempenho preditivo.

Tabela 4.10 – Resultados de desempenho dos classificadores relativos aos métodos de *oversampling* aplicados

Método de *oversampling*	Modelo	ACC	AUC	*F1-Score*	*Precision* - 0	*Precision* - 1	*Recall* - 0	*Recall* - 1
Teste - *Baseline*	ABC	95,8 ± 0,4	98,7 ± 0,3	97,4 ± 0,2	97,1 ± 0,4	90,1 ± 1,1	97,8 ± 0,3	87,9 ± 1,9
	GBT	**96,3 ± 0,5**	99,1 ± 0,2	**97,7 ± 0,3**	97,2 ± 0,4	**92,4 ± 1,8**	**98,2 ± 0,4**	88,2 ± 1,6
	LRC	85,0 ± 0,7	89,9 ± 0,6	91,2 ± 0,4	86,6 ± 0,8	71,1 ± 2,0	96,3 ± 0,3	38,2 ± 4,5
	MLP	93,2 ± 0,3	97,1 ± 0,3	95,8 ± 0,2	94,4 ± 0,4	87,3 ± 1,3	97,3 ± 0,4	76,2 ± 2,1
	RFC	96,2 ± 0,4	**99,2 ± 0,1**	**97,7 ± 0,3**	**98,0 ± 0,4**	89,4 ± 1,5	97,4 ± 0,4	91,7 ± 1,8
	SVM	88,7 ± 0,5	95,0 ± 0,4	92,6 ± 0,3	**98,0 ± 0,2**	64,7 ± 1,2	87,8 ± 0,7	**92,5 ± 0,9**
Validação - *Baseline*	ABC	96,6	98,8	97,9	97,7	92,2	98,2	90,4
	GBT	**96,9**	99,2	**98,1**	97,6	**93,5**	**98,5**	90,2
	LRC	84,7	89,3	91,0	86,3	70,2	96,2	36,8
	MLP	93,6	97,3	96,1	94,8	88,2	97,5	77,7
	RFC	96,7	**99,3**	97,9	**98,3**	90,1	97,5	**93,2**
	SVM	89,4	95,2	93,1	98,2	66,2	88,5	93,3
Teste - Reamostragem Aleatória	ABC	98,7 ± 0,2	99,8 ± 0,1	**98,7 ± 0,2**	**99,6 ± 0,1**	**97,9 ± 0,3**	**97,9 ± 0,3**	**99,6 ± 0,1**
	GBT	**98,6 ± 0,2**	**99,9 ± 0,0**	98,6 ± 0,2	**99,6 ± 0,1**	97,7 ± 0,4	97,7 ± 0,4	**99,6 ± 0,1**
	LRC	86,0 ± 0,5	91,7 ± 0,4	85,9 ± 0,5	86,6 ± 0,7	85,4 ± 0,6	85,2 ± 0,7	86,8 ± 0,8
	MLP	92,8 ± 0,5	98,2 ± 0,2	92,5 ± 0,6	95,8 ± 0,4	90,1 ± 0,7	89,5 ± 0,8	96,1 ± 0,4
	RFC	98,1 ± 0,3	99,8 ± 0,0	98,1 ± 0,3	99,4 ± 0,2	96,8 ± 0,5	96,7 ± 0,5	99,5 ± 0,2
	SVM	91,5 ± 0,4	96,5 ± 0,3	91,2 ± 0,5	94,9 ± 0,4	88,6 ± 0,5	87,8 ± 0,6	95,3 ± 0,4
Validação - Reamostragem Aleatória	ABC	96,4	98,6	97,8	97,8	**92,2**	**98,2**	89,3
	GBT	**96,6**	99,1	**97,9**	97,7	92,0	98,1	90,5
	LRC	85,2	91,3	90,2	96,1	58,1	85,0	85,9
	MLP	90,9	97,3	94,1	98,5	69,6	90,1	94,3
	RFC	96,5	**99,3**	97,8	**98,6**	88,4	97,0	**94,5**
	SVM	89,4	95,3	93,1	98,1	66,3	88,6	92,9
Teste - SMOTE	ABC	97,2 ± 0,2	99,5 ± 0,1	97,2 ± 0,2	97,7 ± 0,4	96,7 ± 0,3	96,6 ± 0,4	97,8 ± 0,4
	GBT	**97,4 ± 0,3**	**99,7 ± 0,0**	**97,4 ± 0,3**	97,8 ± 0,2	**97,0 ± 0,4**	**97,0 ± 0,5**	97,8 ± 0,2
	LRC	86,7 ± 0,5	92,1 ± 0,3	86,5 ± 0,5	87,5 ± 0,4	85,9 ± 0,8	85,6 ± 0,9	87,7 ± 0,5
	MLP	92,7 ± 0,5	98,2 ± 0,2	92,5 ± 0,5	95,6 ± 0,5	90,0 ± 0,6	89,4 ± 0,7	95,9 ± 0,5
	RFC	**97,4 ± 0,3**	**99,7 ± 0,0**	97,3 ± 0,3	**98,3 ± 0,4**	96,4 ± 0,2	96,4 ± 0,3	**98,3 ± 0,4**
	SVM	91,7 ± 0,3	96,5 ± 0,3	91,4 ± 0,3	95,1 ± 0,6	88,8 ± 0,3	87,9 ± 0,4	95,4 ± 0,5
Validação - SMOTE	ABC	96,5	98,8	97,8	98,4	88,9	97,2	93,7
	GBT	**96,6**	**99,3**	**97,9**	98,5	**89,5**	**97,3**	93,8
	LRC	85,6	91,5	90,5	96,2	58,9	85,5	86,1
	MLP	90,5	97,5	93,9	98,4	68,8	89,7	94,0
	RFC	96,3	**99,3**	97,7	**98,8**	87,0	96,6	**95,1**
	SVM	89,4	95,5	93,1	98,1	66,3	88,6	93,0

Fonte: Do Autor (2022).

Quadro 4.4 – Resultados do teste *t* de *Student* para cada par de classificadores treinados com *oversampling* via Reamostragem Aleatória

	ABC	GBT	LRC	MLP	RFC
GBT	0,347				
LRC	0,000	0,000			
MLP	0,000	0,000	0,000		
RFC	0,039	0,078	0,000	0,000	
SVM	0,000	0,000	0,000	0,000	0,000

Fonte: Do Autor (2022).

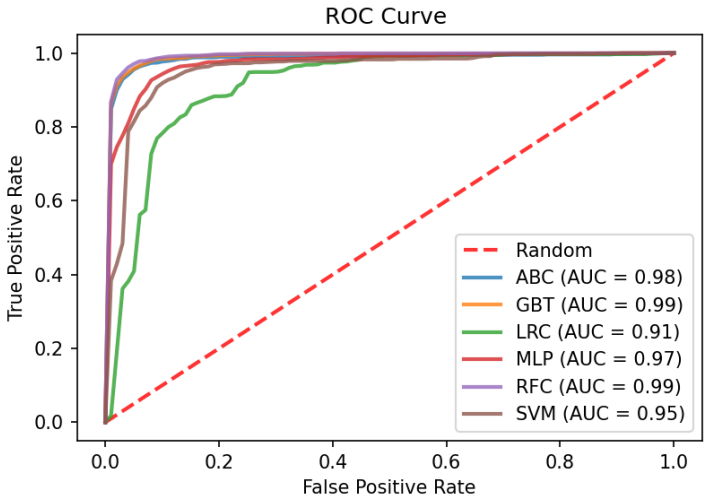
Quadro 4.5 – Resultados do teste *t* de *Student* para cada par de classificadores treinados com *oversampling* via SMOTE

	ABC	GBT	LRC	MLP	RFC
GBT	0,010				
LRC	0,000	0,000			
MLP	0,000	0,000	0,000		
RFC	0,008	0,057	0,000	0,000	
SVM	0,000	0,000	0,000	0,003	0,000

Fonte: Do Autor (2022).

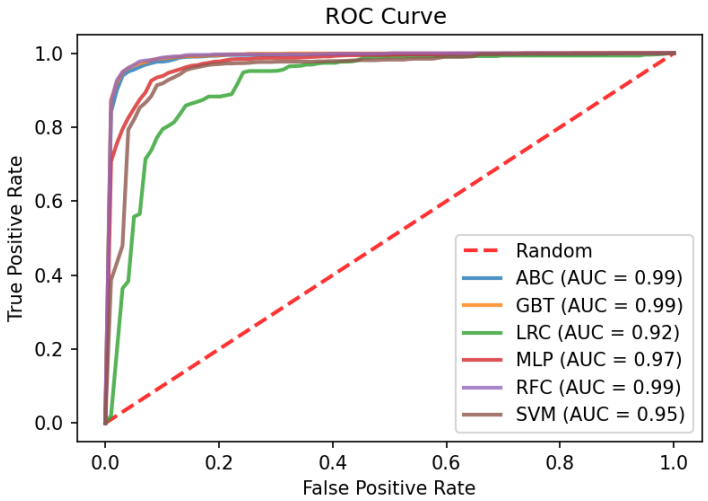
Seguinte aos resultados numéricos, foram gerados os resultados gráficos por meio das Curvas ROC. As Curvas ROC para os classificadores ABC, GBT, LRC, MLP e SVM para a superamostragem por meio da Reamostragem Aleatória e por meio do SMOTE estão contidas nas Figuras 4.10 e 4.11, respectivamente. Observando os resultados gráficos e confrontando-os com os resultados numéricos de classificação obtidos na Tabela 4.10 e nos Quadros 4.4 e 4.5, pode-se verificar que as Curvas ROC corroboram com os resultados numéricos, mostrando que os classificadores ABC, GBT e RFC possuem melhor desempenho que os demais. Além disto, as Curvas ROC dos classificadores de melhor desempenho se sobrepõem, apresentando, graficamente, uma mínima diferença entre o desempenho destes modelos. Realizando um paralelo com as Curvas ROC do treinamento tomado como *baseline*, dispostas na Figura 4.9, observa-se que tal situação também se apresenta para o *baseline*, onde os classificadores ABC, GBT e RFC apresentam os melhores resultados com curvas sobrepostas entre si.

Figura 4.10 – Curva ROC para os classificadores treinados com *oversampling* via Reamostragem Aleatória.



Fonte: Fonte: Do Autor (2022).

Figura 4.11 – Curva ROC para os classificadores treinados com *oversampling* via SMOTE.



Fonte: Fonte: Do Autor (2022).

Por meio das análises dos resultados numéricos e gráficos dispostos nesta seção, o uso de *oversampling* sobre a atual base de dados disponível não implicou em ganho de desempenho preditivo significativo. Desta forma, conforme a inclusão de dados sintéticos não acarreta em ganho de informação, uma vez que os registros incluídos são uma cópia dos cadastros já existentes (para o caso da Reamostragem Aleatória) ou registros sintéticos (para o caso do SMOTE), esta análise indica o não uso de superamostragem na base de dados. Isto se deve ao procedimento de superamostragem gerar um aumento no custo computacional dos classificadores devido ao aumento amostral sem ganho de informação e sem aumento no desempenho preditivo. Portanto, não se justifica o aumento amostral da base de dados via *oversampling*, sendo, assim, os modelos utilizados para as posteriores análises de interpretação serão os classificadores treinados de acordo com o *baseline*, sem o uso de superamostragem.

4.4 Resultados de interpretação dos classificadores

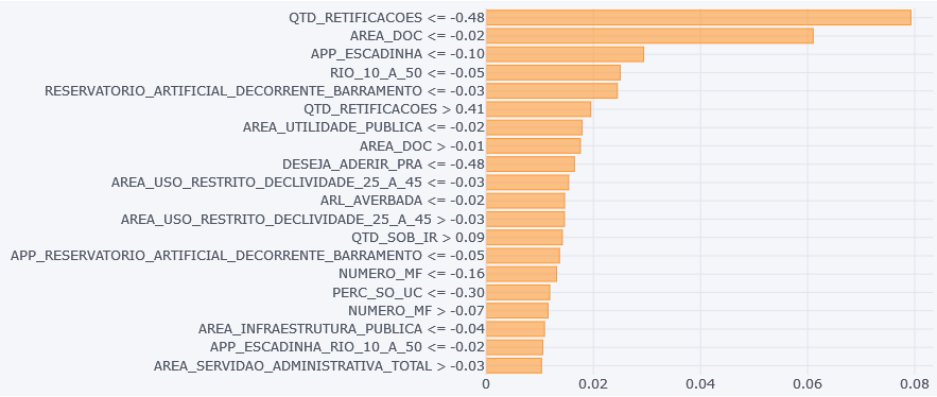
Por meio das análises de classificação realizadas nas seções 4.2 e 4.3, os classificadores apontados como os de melhor desempenho preditivo foram o ABC, o GBT e o RFC, entretanto, uma vez que o LRC possui uma interpretação por meio de seu vetor de pesos, este também terá sua interpretação analisada nesta seção. Desta forma, serão aplicados o LIME, via *Submodular Pick* (SP-LIME), onde os resultados são vistos na subseção 4.4.1. Além disto, serão analisadas as interpretações internas do GBT, LRC e RFC, tendo seus

resultados apresentados na subseção 4.4.2. Estas análises serão realizadas com duas finalidades: elencar quais atributos possuem maior peso nas interpretações e comparar quais interpretações possuem, ou não, correspondência prática, ou seja, se determinada interpretação condiz com as análises realizadas por parte dos especialistas do CAR.

4.4.1 Interpretações geradas pelo LIME

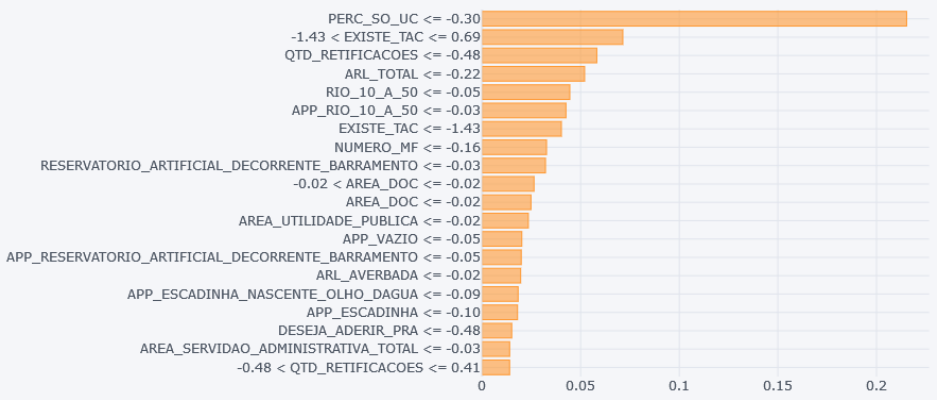
Primeiramente, serão utilizadas as interpretações do LIME por meio da extensão *Submodular Pick*. O parâmetro a ser ajustado do LIME foi a largura do *kernel*, onde menores valores implicam em uma predição local mais acurada, contudo, as medições dos pesos obtidos é demasiadamente baixa (da ordem de 10^{-6}), enquanto um *kernel* mais largo propicia melhores medições dos pesos, todavia, as predições locais possuem maior discrepância das geradas pelos classificadores. Desta forma, a largura do *kernel* (*kw*) foi ajustada manualmente na faixa de valores $kw = [0,5;\ 0,6;\ 0,7;\ 0,8;\ 0,9;\ 1,0;\ 1,1;\ 1,2; 1,3;\ 1,4;\ 1,5]$. A largura do *kernel* precisa ser ajustada para que as interpretações possuam um peso significativo para as visualizações e de maneira que as predições locais possam ser compatíveis com as geradas pelos classificadores. Desta forma, os valores de *kw* obtidos para os classificadores ABC, GBT, LRC e RFC foram 1,3; 1,1; 1,5 e 1,5, respectivamente. Após o ajuste de *kw*, foram geradas 20 interpretações para cada classificador. Os gráficos de interpretação absoluta, ou seja, os gráficos que apresentam o ranqueamento das interpretações de maior impacto para os modelos ABC, GBT, LRC e RFC seguem, respectivamente, nas Figuras 4.12, 4.13, 4.14, 4.15.

Figura 4.12 – Interpretações geradas pelo LIME contendo os pesos absolutos para o ranqueamento das interpretações geradas do classificador ABC.



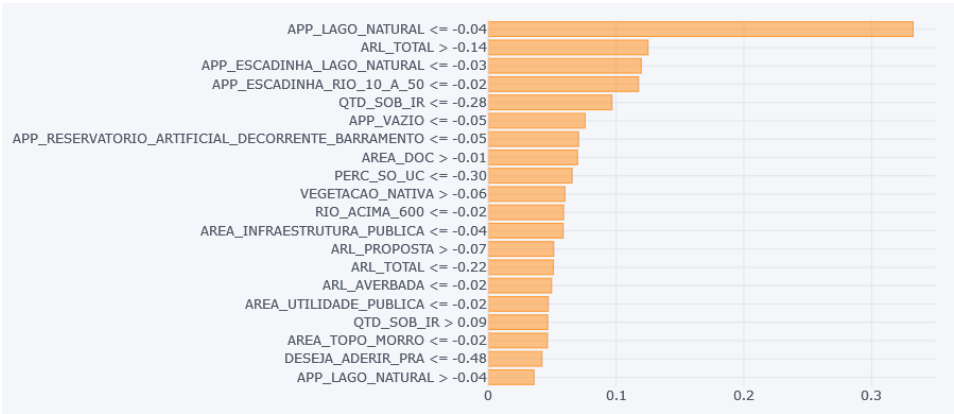
Fonte: Fonte: Do Autor (2022).

Figura 4.13 – Interpretações geradas pelo LIME contendo os pesos absolutos para o ranqueamento das interpretações geradas do classificador GBT.



Fonte: Fonte: Do Autor (2022).

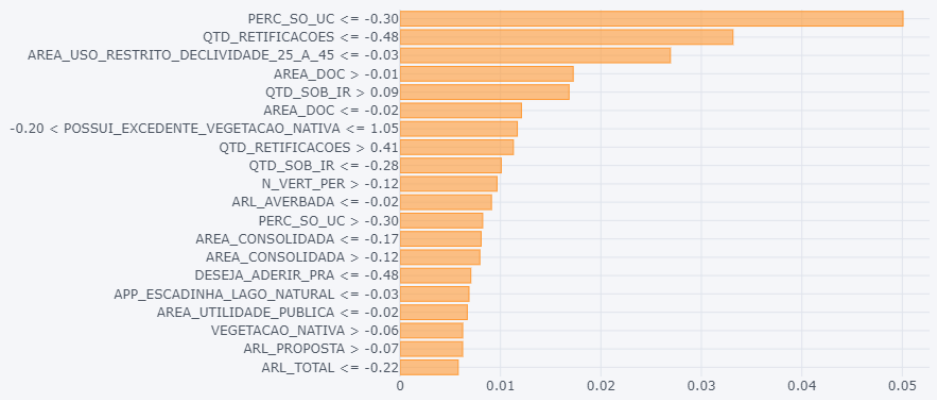
Figura 4.14 – Interpretações geradas pelo LIME contendo os pesos absolutos para o ranqueamento das interpretações geradas do classificador LRC.



Fonte: Fonte: Do Autor (2022).

Analisando o ranqueamento das interpretações absolutas, alguns atributos aparecem em comum para as interpretação de todos os classificadores, sendo estes: ‘AREA DOC’, que se refere à área contida no documento apresentado pelo cadastrante; ‘AREA UTILIDADE PUBLICA’, que corresponde à área, dentro do imóvel rural, ocupada por algum tipo de obra pública; ‘DESEJA ADERIR PRA’, variável categórica onde o cadastrante informa, ou não, se deseja aderir ao Programa de Regularização Ambiental; ‘ARL AVERBADA’, que representa a área de Reserva Legal registrada no imóvel rural; e ‘PERC SO UC’, atributo que trata da sobreposição do imóvel rural com unidades de conservação. Ou seja, as interpretações em comum estão situadas em diferentes grupos de atributos, tendo atributos de área do imóvel rural (‘AREA DOC’), de feições

Figura 4.15 – Interpretações geradas pelo LIME contendo os pesos absolutos para o ranqueamento das interpretações geradas do classificador RFC.



Fonte: Fonte: Do Autor (2022).

do terreno (‘AREA UTILIDADE PUBLICA’, ‘ARL AVERBADA’), de sobreposição (‘PERC SO UC’) e do questionário (‘DESEJA ADERIR PRA’).

Dentre as interpretações geradas pelo ABC, dispostas na Figura 4.12, os cinco atributos que obtiveram maior impacto foram ‘QTD RETIFICACOES’, que se refere à quantidade de retificações realizadas pelo cadastrante no CAR; ‘AREA DOC’, já detalhada no parágrafo anterior, ‘APP ESCADINHA’, que consiste na área de proteção permanente cuja área necessária é computada pela “regra da escadinha”, uma regra que indica o tamanho da área a ser preservada de acordo com a área do imóvel rural; ‘RIO 10 A 50’, no qual é referida a área, dentro do imóvel rural, ocupada por rios de largura de 10 à 50 metros; e ‘RESERVATORIO ARTIFICIAL DECORRENTE BARRAMENTO’, atributo que informa a área de reservatórios de água artificiais decorrentes de barramento ou represamento do curso d’água. Com exceção do primeiro atributo, os demais são todos atributos referentes à área do imóvel rural, sendo ‘AREA DOC’ correspondente à área total da propriedade e os demais atributos de área inseridos no campo das feições do terreno.

As cinco interpretações mais relevantes geradas pelo GBT, vistas na Figura 4.13 com maior impacto nas classificações são ‘PERC SO UC’; ‘EXISTE TAC’, atributo categórico que diz se o cadastro realizado possui, ou não, um termo de ajuste de conduta para a regularização ambiental do imóvel rural (também mencionado na seção 4.2); a ‘QTD RETIFICACOES’; ‘ARL TOTAL’, atributo no qual é inserida a área de reserva legal total da propriedade rural; e a área composta por rios de 10 à 50 metros, dada por ‘RIO 10 A 50’. Observa-se, também, assim como nas interpretações geradas pelo ABC, a variedade dos tipos de atributos entre os de maior relevância, contendo atributos sobre as retificações, sobreposição, questionário e feições do imóvel rural.

As interpretações geradas pelo LRC, localizadas na Figura 4.14 mostram uma maior concentração dos atributos relacionados às feições do imóvel rural dentre as cinco interpretações de maior relevância. Os atributos de maior importância obtidos foram 'APP LAGO NATURAL', que contém a área de mata a ser preservada em volta de espelhos d'água naturais (lagos ou lagoas); a área total da reserva legal, dada por 'ARL TOTAL'; 'APP ESCADINHA LAGO NATURAL', atributo de área de proteção permanente a ser recuperada no entorno de lagos ou lagoas computada pela "regra da escadinha"; 'APP ESCADINHA RIO 10 A 50', atributo relacionado à área de proteção permanente de rios de largura de 10 à 50 metros computada pela "regra da escadinha"; e 'QTD SOB IR' que corresponde à sobreposição do imóvel rural do registro do CAR com outros imóveis rurais.

As cinco interpretações de maior impacto geradas pelo classificador RFC, representadas na Figura 4.15 são a sobreposição sobre unidades de conservações, 'PERC SO UC'; a quantidade de retificações, 'QTD RETIFICACOES'; 'AREA USO RESTRITO DECLIVIDADE 25 A 45', atributo dentro do campo das feições do imóvel rural que corresponde à área onde seu manejo só é permitido acompanhado de órgão ambiental e de maneira sustentável, esta área também possui uma inclinação entre 25 à 45 graus; a área da propriedade registrada pelo documento inserido no CAR, 'AREA DOC'; e a sobreposição do imóvel rural do registro sobre outras propriedades, 'QTD SOB IR'. Os atributos de maior impacto para o RFC são dados por atributos de sobreposição, área total do imóvel e feições da propriedade.

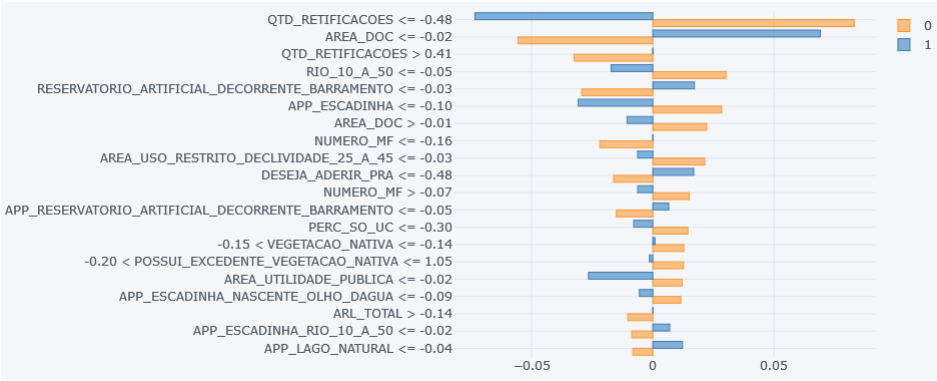
Outro ponto a se destacar são os atributos que mais se apresentaram entre as 5 interpretações de maior impacto, os quais foram 'QTD RETIFICACOES', presente entre as cinco interpretações de maior impacto três vezes e, presentes em duas vezes, os atributos 'AREA DOC', 'PERC SO UC', 'RIO 10 A 50', 'ARL TOTAL' e 'QTD SOB IR'. Além disto, alguns dos atributos presentes nas interpretações obtidas pelo LIME para os quatro classificadores em questão apresentam relevância prática, ou seja, são condizentes com as análises manuais do CAR, realizadas por meio de especialistas. Os atributos de relevância prática são 'QTD RETIFICACOES', 'QTD SOB IR', 'PERC SO UC', 'EXISTE TAC' e 'N VERT PER', esta última referente ao número de vértices do polígono do mapa do imóvel rural inserido do registro do CAR.

As interpretações geradas pelo GBT, vistas na Figura 4.13 mostram uma disposição de variáveis com maior relevância prática com maior ranqueamento, sendo as 3 primeiras ranqueadas 'PERC SO UC', 'EXISTE TAC' e 'QTD RETIFICACOES'. Enquanto o primeiro e o terceiro atributos tiveram suas importâncias descritas no parágrafo anterior, o atributo 'EXITE TAC' infere se o cadastro realizado possui ou não, um termo de ajuste de conduta para a regularização ambiental do imóvel rural (também mencionado na seção 4.2). Assim, um

cadastro que possua um TAC, tende a buscar a regularização ambiental da propriedade e, consequentemente, ter o CAR devidamente registrado e aprovado.

Além das interpretações absolutas, que apresentam um ranqueamento de quais intervalos de predição possuem maior relevância, o SP-LIME também gera uma interpretação por classe, onde o intervalo de predição indica como influencia para cada classe (se contribui para o aumento ou a diminuição da probabilidade do conjunto de dados pertencer a determinada classe). Os gráficos contendo as interpretações do LIME relativas às classes para os classificadores ABC, RFC, LRC e RFC estão dispostos nas Figuras 4.16, 4.17, 4.18 e 4.19, respectivamente. As interpretações por classe permitem uma visão mais detalhada de como cada intervalo de predição contribui, todavia, os intervalos se referem aos atributos normalizados pelo *z-score* que não é uma normalização proporcional como uma mudança de escala (normalização *min-max*). Desta forma, é importante analisar os intervalos com os valores reais dos atributos, de maneira que se possa verificar se o intervalo e a predição gerada possui correspondência por parte da aplicação. Para ampliar as interpretações geradas pelo LIME, as interpretações foram dispostas em tabelas, onde são apresentados os intervalos com seus respectivos valores reais, complementando as análises geradas pelos gráficos. Os resultados de interpretação contendo os valores de cada contribuição dos intervalos, estão contidos, respectivamente nas Tabelas 4.11, 4.12, 4.13 e 4.14.

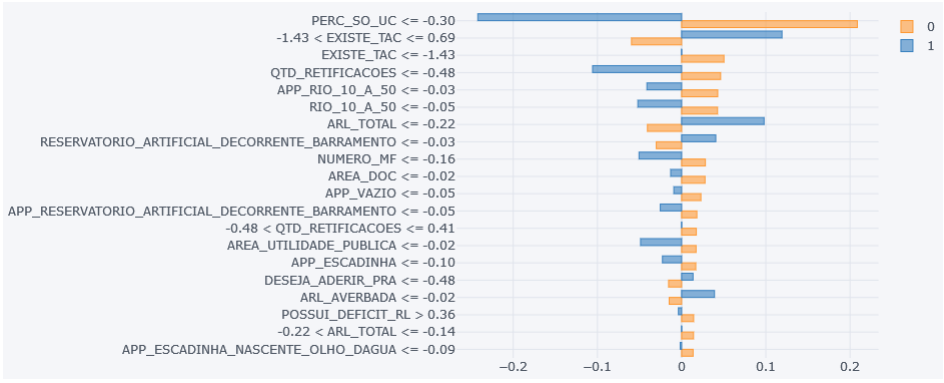
Figura 4.16 – Interpretações geradas pelo LIME contendo os pesos relativos por classe para as interpretações do classificador ABC. Onde ‘0’ se refere à classe de registros Cancelados e ‘1’ se refere à classe de registros Aprovados.



Fonte: Fonte: Do Autor (2022).

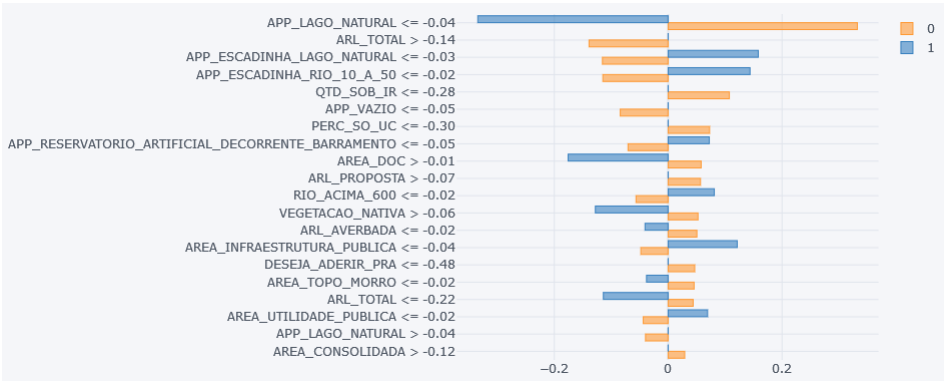
Analisando conjuntamente as interpretações gráficas e numéricas, pontos gerais e específicos para cada predição podem ser elencados. Primeiramente, de uma maneira geral, analisando as Figuras 4.16, 4.17, 4.18 e 4.19, pode-se perceber que os intervalos possuem similaridade entre si, sobretudo, nos valores dos intervalos de predição. Outro ponto geral a ser destacado são as interpretações com inconsistências, como as

Figura 4.17 – Interpretações geradas pelo LIME contendo os pesos relativos por classe para as interpretações do classificador GBT. Onde ‘0’ se refere à classe de registros Cancelados e ‘1’ se refere à classe de registros Aprovados.



Fonte: Fonte: Do Autor (2022).

Figura 4.18 – Interpretações geradas pelo LIME contendo os pesos relativos por classe para as interpretações do classificador LRC. Onde ‘0’ se refere à classe de registros Cancelados e ‘1’ se refere à classe de registros Aprovados.

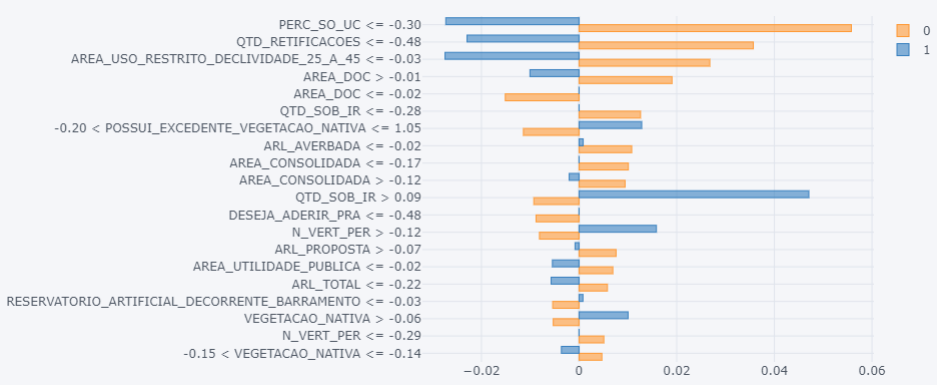


Fonte: Fonte: Do Autor (2022).

áreas negativas e a interpretação do atributo ‘PERC SO UC’, onde, para este último, o valor é incompatível com um valor percentual de sobreposição, além do valor negativo. Em relação às áreas negativas, há duas interpretações possíveis, a primeira considerando o valor como zero, no qual uma interpretação com intervalo maior, ou maior/igual à zero (> 0 ou ≥ 0) poderia ser validada. A segunda possibilidade seria a própria invalidade da interpretação gerada, vista para um intervalo menor ou menor/igual a zero (≤ 0 ou < 0).

Em específico às predições geradas pelo ABC, dispostas pela Figura 4.16 e pela Tabela 4.11, as 5 interpretações de maior relevância foram a ‘QTD RETIFICACOES’, que apresentam dois intervalos: ‘QTD RETIFICACOES’ ≤ 0 (o qual pode ser interpretado somente como igual a zero, por não haver retificações negativas), que apresenta uma maior probabilidade de cancelamento, juntamente com menor probabilidade de

Figura 4.19 – Interpretações geradas pelo LIME contendo os pesos relativos por classe para as interpretações do classificador RFC. Onde ‘0’ se refere à classe de registros Cancelados e ‘1’ se refere à classe de registros Aprovados.



Fonte: Fonte: Do Autor (2022).

Tabela 4.11 – Resultados numéricos das predições do LIME, contendo os pesos de cada predição relativos à cada classe e o peso absoluto para as interpretações geradas pelo classificador ABC. Os valores entre parênteses aos intervalos de predição correspondem aos valores reais do atributo. Os intervalos de predição em negrito correspondem às predições com inconsistência por área negativa.

Intervalo de predição	Cancelados - 0	Aprovados - 1	Peso Absoluto
QTD_RETIFICACOES ≤ -0,48 (0)	0,0832	-0,0734	0,0832
AREA_DOC ≤ -0,02 (-246,021)	-0,0556	0,0691	0,0556
QTD_RETIFICACOES > 0,41 (1)	-0,0325	0,0000	0,0325
RIO_10_A_50 ≤ -0,05 (-0,0197)	0,0302	-0,0172	0,0302
RESERVATORIO_ARTIFICIAL_DECORRENTE_BARRAMENTO ≤ -0,03 (-0,0112)	-0,0293	0,0172	0,0293
APP_ESCADINHA ≤ -0,10 (-0,0199)	0,0284	-0,0308	0,0284
AREA_DOC > -0,01 (499,1165)	0,0222	-0,0107	0,0222
NUMERO_MF ≤ -0,16 (306)	-0,0219	0,0000	0,0219
AREA_USO_RESTRITO_DECLIVIDADE_25_A_45 ≤ -0,03 (0,01350)	0,0215	-0,0063	0,0215
DESEJA_ADERIR_PRA ≤ -0,48 (0)	-0,0162	0,0169	0,0162
NUMERO_MF > -0,07 (44980,65)	0,0151	-0,0062	0,0151
APP_RESERVATORIO_ARTIFICIAL_DECORRENTE_BARRAMENTO ≤ -0,05 (-0,0045)	-0,0150	0,0066	0,0150
PERC_SO_UC ≤ -0,30 (-343)	0,0145	-0,0079	0,0145
-0,15 (3,6739) VEGETACAO_NATIVA ≤ -0,14 (23,2344)	0,0128	0,0009	0,0128
-0,20 (0) POSSUI_EXCEDENTE_VEGETACAO_NATIVA ≤ 1,05 (1)	0,0127	-0,0014	0,0127
AREA_UTILIDADE_PUBLICA ≤ -0,02 (-0,0108)	0,0121	-0,0265	0,0121
APP_ESCADINHA_NASCENTE_OLHO_DAGUA ≤ -0,09 (-0,0001)	0,0116	-0,0055	0,0116
ARL_TOTAL > -0,14 (66,2940)	-0,0103	0,0000	0,0103
APP_ESCADINHA_RIO_10_A_50 ≤ -0,02 (0,0020)	-0,0087	0,0070	0,0087
APP_LAGO_NATURAL ≤ -0,04 (0)	-0,0082	0,0123	0,0082

Fonte: Do Autor (2022).

aprovação, o outro intervalo para as retificações é ‘QTD RETIFICACOES’ $>$ 1, este intervalo indica o oposto do primeiro intervalo, desta maneira, pode-se inferir que um maior número de retificações implica em uma menor probabilidade do registro ser cancelado. Os atributos ‘RIO 10 A 50’ e ‘RESERVATORIO ARTIFICIAL DECORRENTE BARRAMENTO’ apresentaram intervalos inconsistentes por área negativa. A área inserida no documento, ‘AREA DOC’, apresentou 2 intervalos de predição, sendo o primeiro inconsistente por

Tabela 4.12 – Resultados numéricos das predições do LIME, contendo os pesos de cada predição relativos à cada classe e o peso absoluto para as interpretações geradas pelo classificador GBT. Os valores entre parênteses aos intervalos de predição correspondem aos valores reais do atributo. Os intervalos de predição em negrito correspondem às predições com inconsistência por área negativa.

Predição	Cancelados - 0	Aprovados - 1	Peso Absoluto
PERC_SO_UC $\leq$ -0,30 (-343)	0,2086	-0,2419	0,2086
-1,43 (-1) EXISTE_TAC $\leq$ 0,69 (0)	-0,0594	0,1191	0,0594
EXISTE_TAC $\leq$ -1,43 (-1)	0,0503	0,0000	0,0503
QTD_RETIFICACOES $\leq$ -0,48 (0)	0,0463	-0,1057	0,0463
APP_RIO_10_A_50 $\leq$ -0,03 (0,0032)	0,0429	-0,0412	0,0429
RIO_10_A_50 $\leq$ -0,05 (-0,0197)	0,0427	-0,0519	0,0427
ARL_TOTAL $\leq$ -0,22 (1,7648)	-0,0405	0,0979	0,0405
RESERVATORIO_ARTIFICIAL_DECORRENTE_BARRAMENTO $\leq$ -0,03 (-0,0112)	-0,0301	0,0406	0,0301
NUMERO_MF $\leq$ -0,16 (306)	0,0282	-0,0505	0,0282
AREA_DOC $\leq$ -0,02 (-246,021)	0,0278	-0,0130	0,0278
APP_VAZIO $\leq$ -0,05 (0,0551)	0,0229	-0,0092	0,0229
APP_RESERVATORIO_ARTIFICIAL_DECORRENTE_BARRAMENTO $\leq$ -0,05 (-0,0045)	0,0181	-0,0253	0,0181
-0,48 (0) QTD_RETIFICACOES $\leq$ 0,41 (1)	0,0175	0,0000	0,0175
AREA_UTILIDADE_PUBLICA $\leq$ -0,02 (-0,0108)	0,0172	-0,0487	0,0172
APP_ESCADINHA $\leq$ -0,10 (-0,0199)	0,0168	-0,0227	0,0168
DESEJA_ADERIR_PRA $\leq$ -0,48 (0)	-0,0154	0,0136	0,0154
ARL_AVERBADA $\leq$ -0,02 (0,4159)	-0,0147	0,0390	0,0147
POSSUI_DEFICIT_RL $>$ 0,36 (0)	0,0143	-0,0038	0,0143
-0,22 (1,7648) ARL_TOTAL $\leq$ -0,14 (66,2940)	0,0142	0,0000	0,0142
APP_ESCADINHA_NASCENTE_OLHO_DAGUA $\leq$ -0,09 (-0,0001)	0,0136	-0,0017	0,0136

Fonte: Do Autor (2022).

Tabela 4.13 – Resultados numéricos das predições do LIME, contendo os pesos de cada predição relativos à cada classe e o peso absoluto para as interpretações geradas pelo classificador LRC. Os valores entre parênteses aos intervalos de predição correspondem aos valores reais do atributo. Os intervalos de predição em negrito correspondem às predições com inconsistência por área negativa.

Predição	Cancelados - 0	Aprovados - 1	Peso Absoluto
APP_LAGO_NATURAL $\leq$ -0,04 (0)	0,3323	-0,3342	0,3323
ARL_TOTAL $>$ -0,14 (66,2940)	-0,1389	0,0000	0,1389
APP_ESCADINHA_LAGO_NATURAL $\leq$ -0,03 (-0,0044)	-0,1154	0,1583	0,1154
APP_ESCADINHA_RIO_10_A_50 $\leq$ -0,02 (0,0020)	-0,1147	0,1438	0,1147
QTD_SOB_IR $\leq$ -0,28 (0,0165)	0,1075	0,0000	0,1075
APP_VAZIO $\leq$ -0,05 (0,0551)	-0,0841	0,0000	0,0841
PERC_SO_UC $\leq$ -0,30 (-343)	0,0728	0,0000	0,0728
APP_RESERVATORIO_ARTIFICIAL_DECORRENTE_BARRAMENTO $\leq$ -0,05 (-0,0045)	-0,0704	0,0723	0,0704
AREA_DOC $>$ -0,01 (499,1165)	0,0580	-0,1756	0,0580
ARL_PROPOSTA $>$ -0,07 (79,4278)	0,0568	0,0000	0,0568
RIO_ACIMA_600 $\leq$ -0,02 (0,2252)	-0,0564	0,0811	0,0564
VEGETACAO_NATIVA $>$ -0,06 (179,7185)	0,0524	-0,1278	0,0524
ARL_AVERBADA $\leq$ -0,02 (0,4159)	0,0505	-0,0406	0,0505
AREA_INFRAESTRUTURA_PUBLICA $\leq$ -0,04 (-0,0012)	-0,0479	0,1213	0,0479
DESEJA_ADERIR_PRA $\leq$ -0,48 (0)	0,0467	0,0000	0,0467
AREA_TOPO_MORRO $\leq$ -0,02 (0,0040)	0,0455	-0,0379	0,0455
ARL_TOTAL $\leq$ -0,22 (1,7648)	0,0439	-0,1138	0,0439
AREA_UTILIDADE_PUBLICA $\leq$ -0,02 (-0,0108)	-0,0437	0,0692	0,0437
APP_LAGO_NATURAL $>$ -0,04 (0)	-0,0397	0,0000	0,0397
AREA_CONSOLIDADA $>$ -0,12 (55,6084)	0,0291	0,0000	0,0291

Fonte: Do Autor (2022).

Tabela 4.14 – Resultados numéricos das predições do LIME, contendo os pesos de cada predição relativos à cada classe e o peso absoluto para as interpretações geradas pelo classificador RFC. Os valores entre parênteses aos intervalos de predição correspondem aos valores reais do atributo. Os intervalos de predição em negrito correspondem às predições com inconsistência por área negativa.

Predição	Cancelados - 0	Aprovados - 1	Peso Absoluto
PERC_SO_UC $\leq$ -0,30 (-343)	0,0558	-0,0273	0,0558
QTD_RETIFICACOES $\leq$ -0,48 (0)	0,0357	-0,0229	0,0357
AREA_USO_RESTRITO_DECLIVIDADE_25_A_45 $\leq$ -0,03 (0,01350)	0,0268	-0,0274	0,0268
AREA_DOC $>$ -0,01 (499,1165)	0,0190	-0,0101	0,0190
AREA_DOC $\leq$ -0,02 (-246,021)	-0,0151	0,0000	0,0151
QTD_SOB_IR $\leq$ -0,28 (0,0165)	0,0126	0,0000	0,0126
-0,20 (0) POSSUI_EXCEDENTE_VEGETACAO_NATIVA $\leq$ 1,05 (1)	-0,0114	0,0128	0,0114
ARL_AVERBADA $\leq$ -0,02 (0,4159)	0,0108	0,0008	0,0108
AREA_CONSOLIDADA $\leq$ -0,17 (3,8955)	0,0101	0,0000	0,0101
AREA_CONSOLIDADA $>$ -0,12 (55,6084)	0,0094	-0,0020	0,0094
QTD_SOB_IR $>$ 0,09 (1,9836)	-0,0093	0,0471	0,0093
DESEJA_ADERIR_PRA $\leq$ -0,48 (0)	-0,0088	0,0000	0,0088
N_VERT_PER $>$ -0,12 (49)	-0,0081	0,0158	0,0081
ARL_PROPOSTA $>$ -0,07 (79,4278)	0,0076	-0,0008	0,0076
AREA_UTILIDADE_PUBLICA $\leq$ -0,02 (-0,0108)	0,0069	-0,0055	0,0069
ARL_TOTAL $\leq$ -0,22 (1,7648)	0,0058	-0,0057	0,0058
RESERVATORIO_ARTIFICIAL_DECORRENTE_BARRAMENTO $\leq$ -0,03 (-0,0112)	-0,0054	0,0008	0,0054
VEGETACAO_NATIVA $>$ -0,06 (179,7185)	-0,0053	0,0100	0,0053
N_VERT_PER $\leq$ -0,29 (6)	0,0051	0,0000	0,0051
-0,15 (3,6739) VEGETACAO_NATIVA $\leq$ -0,14 (23,2344)	0,0047	-0,0036	0,0047

Fonte: Do Autor (2022).

intervalo negativo (‘AREA DOC’ $<$ 0) e o segundo contendo validade (‘AREA DOC’ $>$ 499ha), o qual aponta para um aumento na probabilidade de cancelamento em conjunto com uma redução na probabilidade de aprovação. Outro fator a ser levantando são interpretações geradas de maneira discriminativa, onde os pesos por classe, possuem sinais opostos, visto que, uma interpretação com os pesos por classe de mesmo sinal pode gerar ambiguidades (mesma forma de contribuição para duas classes mutualmente exclusivas).

Sobre as cinco interpretações de maior peso geradas pelo GBT, apresentadas pela Figura 4.17 e pela Tabela 4.12, a primeira interpretação, dada pelo atributo ‘PERC SO UC’ contém uma inconsistência em seu intervalo, conforme supracitado em parágrafos anteriores. O segundo e terceiro intervalos de predição são compostos pelo atributo ‘EXISTE TAC’, que possui somente 3 valores possíveis, (-1, 0 e 1), o primeiro intervalo (-1 <‘EXISTE TAC’ $\leq$ 0), para este caso, conforme o intervalo exclui o -1 e inclui somente o valor de 0, pode-se atribuir este intervalo como ‘EXISTE TAC’ = 0. Sendo assim, para este intervalo, é indicado que os registros do CAR que não informam a presença do Termo de Ajuste de Conduta possuem maiores chances de aprovação e, consequentemente, menores chances de cancelamento. O segundo intervalo para o atributo ‘EXISTE TAC’ é o intervalo ‘EXISTE TAC’ $\leq$ -1, o qual pode ser entendido por ‘EXISTE TAC’ $=$ -1, esta interpretação indica que registros do CAR que não possuem o TAC, tendem a serem cancelados. A

interpretação para o intervalo do atributo ‘QTD RETIFICACOES’ possui o mesmo perfil que a interpretação gerada pelo ABC. O intervalo de predição gerado pelo atributo ‘APP RIO 10 A 50’, dado por ‘APP RIO 10 A 50 $\leq$ 0,0032ha, indica que imóveis rurais com pequenas áreas de proteção permanente para rios de largura de 10 a 50 metros possuem maior probabilidade de cancelamento e menor probabilidade de aprovação. Assim como nos intervalos dados pelo ABC, os intervalos fornecidos pelo GBT possuem baixa ambiguidade geradas pelos sinais dos pesos das interpretações.

Os intervalos gerados pelo LRC, representadas pela Figura 4.18 e pela Tabela 4.13, possuem, assim como os dois primeiros, uma baixa ambiguidade das interpretações devido ao mesmo sinal para uma dada predição. Acerca das 5 interpretações de maior peso, a interpretação dada pelo atributo ‘APP LAGO NATURAL’, por meio do intervalo APP LAGO NATURAL $\leq$ 0ha, o qual pode ser aplicado como APP LAGO NATURAL $=$ 0ha, por se tratar de um atributo de área, implica que imóveis rurais que não possuem área para esta APP possuem maiores chances de cancelamento. O intervalo de predição da *feature* ‘ARL TOTAL’ $>$ 66,3ha, no qual se pode inferir que grandes áreas de reserva legal total indicam maiores probabilidades de cancelamento. O intervalo dado pelo atributo ‘APP ESCADINHA LAGO NATURAL’ apresentou inconsistência por área negativa. A predição intervalar para o atributo ‘APP ESCADINHA RIO 10 A 50’ $\leq$ 0,0020ha, indica que pequenas áreas deste tipo de APP possuem maior probabilidade de ter o registro do CAR aprovado. O intervalo relativo à sobreposição entre imóveis rurais, mostrado pela predição ‘QTD SOB IR’ $\leq$ 0,0165ha, indica que um imóvel rural com baixa sobreposição com outros imóveis rurais possui maior probabilidade de cancelamento.

Com relação às interpretações do LIME geradas para o RFC, vistas na Figura 4.19 e na Tabela 4.14, a interpretação gerada pelo intervalo dado pelo atributo ‘ARL AVERBADA’ $\leq$ 0,41ha possui uma ambiguidade devido ao sinal dos pesos da predição, uma vez que ambos os pesos possuem sinal positivo, gerando uma análise inconclusiva acerca de qual classe tal intervalo possui tendência. Sobre os cinco intervalos de predição mais relevantes, os intervalos ‘PERC SO UC’ $<$ -343 e ‘AREA DOC’ $<$ -246 possuem inconsistências, conforme mencionado anteriormente. O intervalo de predição. Os intervalos de predição ‘QTD RETIFICACOES’ $\leq$ 0 e ‘AREA DOC’ $>$ 499ha possuem o mesmo perfil que os mesmos intervalos de predição obtidos para o classificador ABC. O intervalo de predição dado pela área de uso restrito, mostrado por ‘AREA USO RESTRITO DECLIVIDADE 25 A 45’ $\leq 0,0135$ha mostra que pequenas áreas de uso restrito do tipo de área com inclinação entre 25 e 45 graus possuem maior probabilidade de cancelamento do registro do CAR.

Dentre os intervalos de predição, contabilizando somente os intervalos sem inconsistências, somente o intervalo ‘DESEJA ADERIR PRA’ ≤ 0 foi comum às interpretações geradas por todos os classificadores. Outras interpretações válidas, presentes em três dos quatro conjuntos de interpretações foram: ‘QTD RETIFICACOES’ ≤ 0, ‘AREA DOC’ $>$ 499ha, ‘ARL TOTAL’ $\leq 1,76$ha e ARL AVERBADA $\leq 0,41$ha. Dentre os 5 intervalos de predição de maior importância, o intervalo mais comum foi ‘QTD RETIFICACOES’ ≤ 0 presente em 3 vezes, os demais intervalos válidos não possuíram repetição. Além dos intervalos de maior importância e dos mais presentes, alguns intervalos de predição possuem correspondência com as análises manuais do CAR, sendo estes ‘EXISTE TAC’ $\leq$ -1; ‘QTD RETIFICACOES’ ≤ 0; ‘QTD RETIFICACOES’ > 1; ‘N VERT PER’ > 12; e ‘N VERT PER’ ≤ 6. Os dois últimos se referem ao número de vértices do polígono do terreno do imóvel rural. Seu perfil de predição é condizente com a prática, dado que, durante análises manuais, foram vistos registros irregulares onde apresentaram polígonos com baixo número de vértices, sobretudo polígonos de formato triangular. Uma interpretação com perfil oposto ao visto nas análises manuais foi observada no intervalo de predição ‘QTD SOB IR’ $\leq$ 0,0165ha, no qual as baixas sobreposições que influenciariam na redução da probabilidade de cancelamento, o que ocorre, na prática, de maneira oposta ao apresentado pela interpretação do LIME.

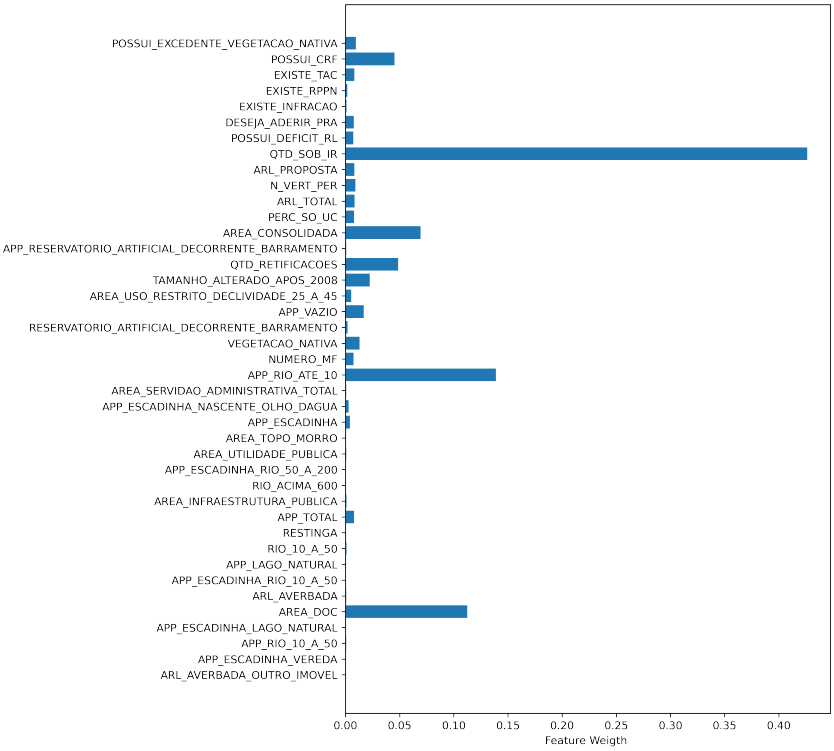
As análises geradas pelo LIME possuem como vantagem a disposição intervalar dos atributos, mostrando a margem de valores nos quais seu intervalo contribui para o aumento da probabilidade de uma classe ou de outra. Entretanto, tais intervalos possuem a ressalva de que nem sempre o intervalo pode fazer correspondência prática. Para verificar isto, se faz necessária a visualização dos intervalos com os valores absolutos dos atributos, sem a normalização aplicada para verificar se determinado intervalo é condizente, ou não, com a prática e se o sinal do peso disposto é correspondente à aplicação final. Portanto, as análises intervalares do LIME requerem uma análise mais detalhada e criteriosa para que se possa extrair melhor os resultados das predições intervalares geradas.

4.4.2 Interpretações obtidas pelos pesos internos dos classificadores

Após os ensaios de interpretação dos classificadores por meio do LIME, um algoritmo de aprendizagem de máquina interpretável, foram geradas as interpretações internas dos classificadores que possuem a propriedade de interpretação por meio de seus pesos. Dentre os modelos de classificação utilizados, o GBT, LRC e o RFC permitem tal interpretação. As interpretações internas geradas permitem analisar as variáveis de maior importância para a classificação, de acordo com o valor do peso obtido para cada atributo. Os gráficos

contendo as interpretações internas para os modelos GBT, LRC e RFC estão dispostos, respectivamente, nas Figuras 4.20, 4.21 e 4.22.

Figura 4.20 – Interpretações internas geradas pelos pesos do classificador GBT.

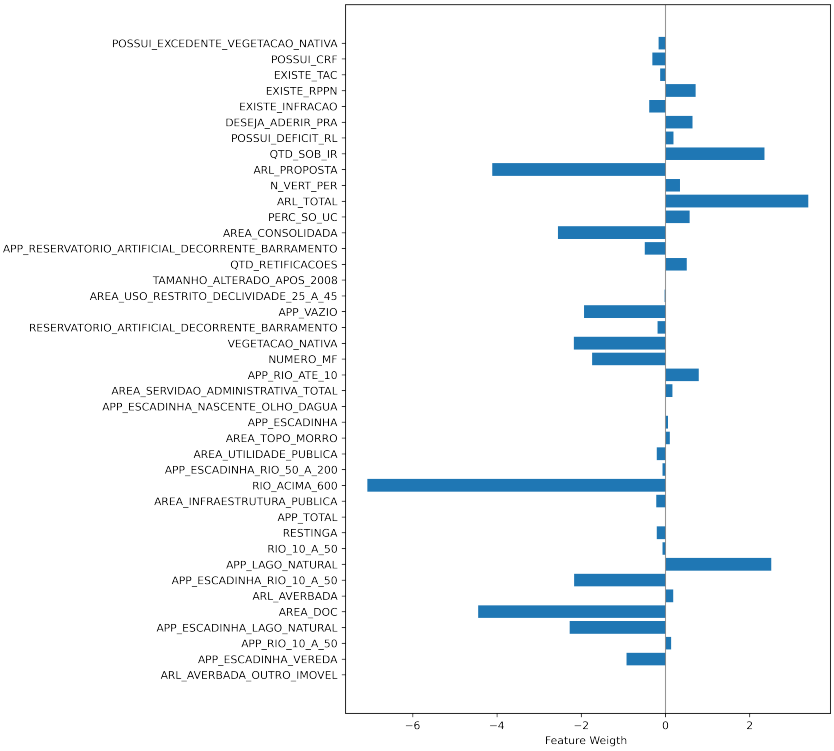


Fonte: Fonte: Do Autor (2022).

No ranqueamento dos atributos realizado pelo GBT (Figura 4.20), observa-se que todos os atributos possuem contribuição positiva. Assim, pode-se inferir que todos os atributos contribuem, em maior ou menor grau, para um aumento da probabilidade de aprovação dos registros. Conforme não foram apresentados pesos negativos, o enfoque será dado somente no ranqueamento. Os 5 atributos de maior contribuição foram ‘QTD SOB IR’ que corresponde à sobreposição do imóvel rural do registro do CAR com outros imóveis rurais; ‘APP RIO ATE 10’, que representa a área, dentro do imóvel rural, ocupada por rios de largura do curso de até 10 metros; ‘AREA DOC’, que se refere à área contida no documento apresentado pelo cadastrante; ‘AREA CONSOLIDADA’, que representa a área com ocupação humana (lavouras, pastagens, etc.) desde antes de 22 de julho de 2008; e ‘POSSUI CRF’, atributo categórico de resposta que indica se o imóvel rural possui, ou não, cota de reserva florestal (CRF), que corresponde por área de vegetação nativa, servidão florestal, Reserva Particular do Patrimônio Natural (RPPN) ou reserva legal. Os atributos destacados consistem em

atributos de área total do imóvel (‘AREA DOC’), feições do imóvel rural (‘APP RIO ATE 10’ e ‘AREA CONSOLIDADA’), sobreposição (‘QTD SOB IR’) e questionário (‘POSSUI CRF’).

Figura 4.21 – Interpretações internas geradas pelos pesos do classificador LRC.



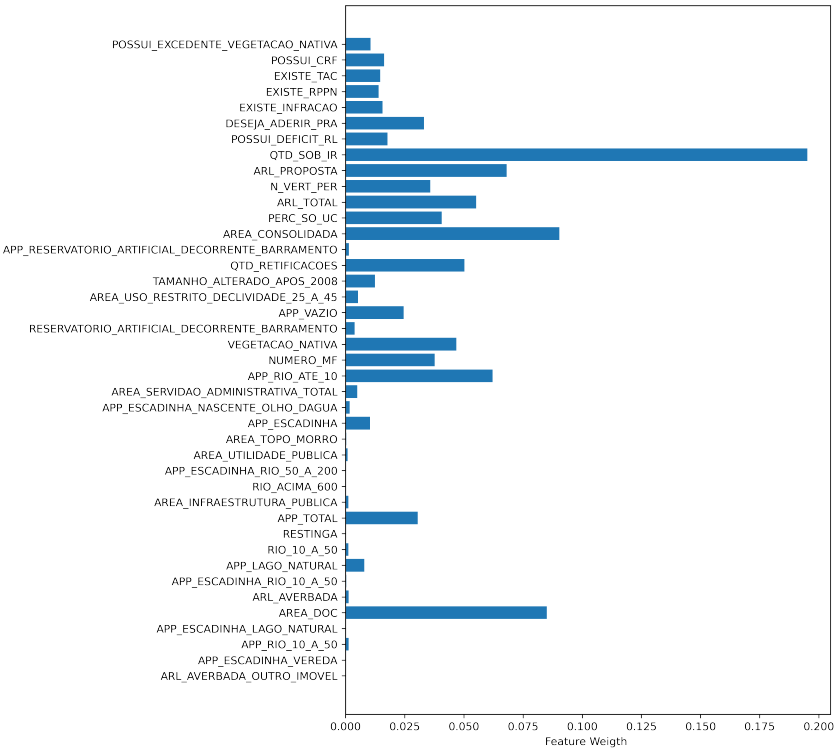
Fonte: Fonte: Do Autor (2022).

As interpretações geradas pelo LRC, vistas na Figura 4.21, apresentam tanto pesos positivos, quanto negativos, o que implica que há atributos que contribuem para o aumento e atributos que contribuem para a redução da probabilidade de aprovação. Os 5 atributos que mais contribuíram para o aumento da probabilidade de aprovação do cadastro, segundo a interpretação gerada, foram: ‘ARL TOTAL’, atributo no qual é inserida a área de reserva legal total da propriedade rural; ‘APP LAGO NATURAL’, que contém a área de mata a ser preservada em volta de espelhos d’água naturais (lagos ou lagoas); ‘QTD SOB IR’ e ‘APP RIO ATE 10’, assim como na interpretação gerada pelo GBT; e ‘EXISTE RPPN’ que consiste em uma variável de resposta do questionário se o imóvel rural possui alguma RPPN. As *features* que correspondem as 5 interpretações de maior peso pertencem aos grupos de feições do imóvel rural (‘ARL TOTAL’, ‘APP LAGO NATURAL’ e ‘APP RIO ATE 10’), sobreposição (‘QTD SOB IR’) e questionário (‘EXISTE RPPN’).

As interpretações de maior peso negativo para o classificador LRC são: ‘RIO ACIMA 600’, atributo que informa a área do imóvel rural coberta por rios de largura do curso acima de 600 metros; ‘AREA DOC’;

‘ARL PROPOSTA’, que representa a área da reserva legal proposta pelo proprietário; ‘AREA CONSOLIDADA’; e ‘APP ESCADINHA LAGO NATURAL’, atributo de área de proteção permanente a ser recuperada no entorno de lagos ou lagoas computada pela “regra da escadinha”.

Figura 4.22 – Interpretações internas geradas pelos pesos do classificador RFC.



Fonte: Fonte: Do Autor (2022).

Sobre as interpretações obtidas pelo RFC, dispostas na Figura 4.22, o ranqueamento obtido foi similar ao ranqueamento gerado pelo GBT, tanto no sentido de não haver atributos com peso negativo, quanto na ordem do ranqueamento dos atributos. Os atributos de interpretação de maior peso, foram ‘QTD SOB IR’; ‘AREA CONSOLIDADA’; ‘AREA DOC’; ‘ARL PROPOSTA’ e ‘APP RIO ATE 10’. Os atributos que obtiveram as 5 interpretações de maior peso pertencem aos grupos de feições do imóvel rural (‘ARL TOTAL’, ‘APP RIO ATE 10’, ‘AREA CONSOLIDADA’), sobreposição (‘QTD SOB IR’) e área total do imóvel rural (‘AREA DOC’).

Comparando os 5 atributos de maior peso, tanto positivo, quanto negativo, observa-se que 4 dos 5 atributos aparecem dispostos em todas as interpretações geradas: ‘AREA DOC’; ‘QTD SOB IR’; ‘ARL PROPOSTA’; e ‘APP RIO ATE 10’. Comparou-se juntamente com os pesos negativos do LRC dado que os maiores pesos negativos também são relevantes para as análises, uma vez que sua influência é dada pelo

valor, o sinal indica para qual classe tende, para um caso de classificação binária. Realizando um comparativo com as interpretações absolutas feitas pelo LIME, os atributos 'AREA DOC' e 'QTD SOB IR' se apresentam presentes em ambas as interpretações dentre as interpretações de maior relevância. Dentre as 5 interpretações de maior relevância, a do atributo 'QTD SOB IR' é uma que possui grande importância prática para as análises manuais do CAR.

As interpretações geradas pelos pesos internos dos classificadores apresentam um ranqueamento geral dos atributos contendo os pesos e seus sinais que indicam como o peso contribui para o aumento ou redução da probabilidade de determinado conjunto de dados pertencer a uma determinada classe. Diferentemente do LIME, onde as interpretações são intervalares, contendo cada atributo dentro de uma faixa de valores, os pesos internos dos modelos apresentam somente os atributos como um todo. Tal fator pode levar a uma vantagem de robustez em relação a intervalos inconsistentes, como visto em algumas predições geradas pelo LIME quando os intervalos eram dispostos em seus valores reais. Todavia, uma não disposição sobre como cada *feature* contribui para o aumento ou redução do aumento da probabilidade de um determinado conjunto é uma lacuna para a interpretação utilizando os pesos dos classificadores. Dado que, sem uma análise intervalar, não é possível inferir se determinado atributo contribui uniformemente, independentemente de seu valor.

Analisando as interpretações geradas pelo LIME e pelos pesos internos dos classificadores que permitem tal análise, pode-se elencar alguns pontos em relação ao uso dos interpretadores para este trabalho. Para uma análise de interpretação local, amostra por amostra, o uso do LIME se faz interessante, dado que o LIME permite tal recurso. Assim, para uma geração de interpretação de como o classificador gerou tal resultado para um único registro, a interpretação do LIME se justifica. Outro ponto a ser destacado, é para o caso de interpretações onde seja necessário gerar uma resposta com uma análise onde o valor dos atributos se faz importante. Para estes casos, o LIME também se sobressai dada a capacidade de geração de interpretações intervalares, no qual o peso por classe é definido para uma faixa específica de valores de cada atributo.

Caso o valor das *features* em relação à interpretação por classe seja irrelevante, o uso somente das interpretações internas dos classificadores é suficiente, uma vez que não será necessário uma implementação de ferramenta adicional. Vale ressaltar que as interpretações intervalares do LIME necessitam ter seus intervalos retornados ao valor real da variável, em casos de uso de normalização de *features*. Tal conversão se faz necessária para fins de compreensão das interpretações geradas e verificação de eventuais incoerências nos intervalos. Como, por exemplo, valores negativos para atributos que representam área, conforme observado neste trabalho. Por fim, o uso de uma ferramenta específica dependerá de qual formato de interpretação se

deseja, devendo ser escolhida por parte da aplicação e sendo feito os devidos procedimentos para compreensão das interpretações geradas.

5 CONCLUSÕES, PERSPECTIVAS E PRÓXIMOS PASSOS

Este trabalho tem como objetivo propor uma aplicação de algoritmos de aprendizagem de máquina na identificação de registros irregulares (espúrios) do CAR, por meio de algoritmos de classificação de dados, onde os registros do CAR são classificados em Aprovados ou Cancelados. Além da classificação, foi aplicado um método de interpretação dos atributos de maior impacto, seja por meio de uma interpretação de relevância absoluta, envolvendo o atributo como um todo, ou seja por uma interpretação intervalar de cada *feature*.

O trabalho seguiu todo um roteiro de procedimentos de mineração de dados, desde a aquisição dos registros, passando pelos processos de limpeza e organização dos cadastros, codificação de campos textuais, onde houve necessidade, seleção de atributos, *oversampling*, aplicação dos classificadores, avaliação dos resultados de classificação e interpretação dos classificadores. Além do *pipeline* padrão, foi estudado o uso de *oversampling* para correção do desbalanceamento e foram avaliados os impactos da utilização de tal prática.

Os resultados exploratórios apontaram inviabilidade do uso dos registros pendentes (não rotulados) para a classificação, sendo necessário, estudos adicionais caso seja desejada a inclusão de tais registros. Durante os ensaios comparativos acerca da seleção de atributos, foi escolhido o uso do Discriminante de Fisher (FDR) com o acréscimo da remoção por correlação de Pearson devido à este método utilizar menos *features* para a classificação. Os resultados de classificação utilizando os métodos de *oversampling* não geraram impacto significativo de melhoria na classificação dos registros, portanto, não sendo necessária a correção do desbalanceamento. Tal fato se deve às atualizações na base de dados, ou seja, com o aumento amostral obtido ao longo do trabalho desta dissertação, os resultados apresentaram melhoria. Em estudos realizados previamente, os resultados preliminares apontavam uma classificação tendenciosa para a classe de Cancelados e grande discrepância entre os resultados, sendo a classe menos numerosa prejudicada significativamente. Quanto aos resultados de classificação, os modelos *AdaBoost* (ABC), *Gradient Boosting* (GBT) e *Random Forest* (RFC) obtiveram os melhores resultados sem diferença estatística significativa entre estes, logo, a escolha de um dos modelos poderá ser feita baseando-se em critérios adicionais como tempo de processamento e capacidade de interpretação interna do classificador. Os resultados de predição atingiram, em todos os índices, mais de 90% para o conjunto de validação, que corresponde aos registros novos aos classificadores. Dentre os 3 modelos, o mais indicado para aplicação foi o *Random Forest* devido ao menor custo computacional e à sua capacidade de interpretação.

Os resultados de interpretação elencaram os atributos de maior importância para a classificação automática dos registros do CAR. Dentre as interpretações observadas, houve a presença de atributos que não

eram mencionadas como relevantes nas análises manuais, como a área presente no documento do imóvel rural, área total da reserva legal, áreas ocupadas por rios e outros tipos de atributos do grupo das feições do imóvel rural. Além da observação de interpretações novas, foram vistas interpretações de atributos já conhecidos das análises manuais via especialistas, como a quantidade de retificações e atributos de sobreposição. Tal situação, além de apontar novas descobertas acerca do comportamento da classificação automática, corrobora com o procedimento consolidado manualmente. Além disto, foram realizadas análises de interpretação dos pesos dos classificadores que possuem interpretabilidade por meio de seus parâmetros. Outro apontamento gerado a partir das interpretações obtidas, é uma possível divisão dos grupos de atributos para uma futura abordagem de classificação de *ensemble* de classificadores, no qual cada classificador responderia por um grupo de atributos.

Para trabalhos futuros, tem-se como propostas: (i) a inclusão de estudos adicionais dos registros pendentes para fins de uma verificação mais detalhada da viabilidade de seu uso ou não; (ii) atualizar os classificadores, de acordo com novos registros sendo aprovados ou cancelados; (iii) usar o conjunto de interpretações globais e locais (amostra por amostra) para verificação e validação prática das interpretações obtidas neste trabalho; (iv) estudos do uso de modelos de *ensemble* para classificação, utilizando modelos especialistas para cada grupo de atributos; e (v) aplicação do sistema completo no sistema do Cadastro Ambiental Rural.

Por fim, os resultados apresentados e discutidos por este trabalho indicam uma viabilidade de implementação de um sistema automático de classificação para os registros do Cadastro Ambiental Rural, utilizando a base de dados final proposta e os modelos de melhor desempenho preditivo elencados. Assim, com futuras atualizações na base de dados e melhorias na própria classificação em si, o método proposto poderá beneficiar as análises do CAR, fornecendo uma análise rápida, detalhada e eficaz, auxiliando especialistas e agricultores do Brasil.

REFERÊNCIAS

AGARWAL, N.; DAS, S. Interpretable machine learning tools: A survey. In: IEEE. **2020 IEEE Symposium Series on Computational Intelligence (SSCI)**. Canberra, Australia, 2020. p. 1528–1534.

AGGARWAL, C. C. **Neural networks and deep learning**: A textbook. 1. ed. Cham, Switzerland: Springer, 2018. 497 p.

AGGARWAL, C. C. et al. **Data mining: the textbook**. New York, USA: Springer, 2015. v. 1.

ALSIRHANI, A.; SAMPALLI, S.; BODORIK, P. Ddos detection system: utilizing gradient boosting algorithm and apache spark. In: IEEE. **2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)**. [S.l.], 2018. p. 1–6.

ARVOR, D. et al. 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**, Elsevier, v. 103, p. 105281, 2021.

AZODI, C. B.; TANG, J.; SHIU, S.-H. Opening the black box: Interpretable machine learning for geneticists. **Trends in Genetics**, Elsevier, v. 36, n. 6, p. 442–455, 2020.

BARBOSA, T. S. et al. Fault detection and classification in cantilever beams through vibration signal analysis and higher-order statistics. **Journal of Control, Automation and Electrical Systems**, Springer, v. 27, n. 5, p. 535–541, 2016.

BRANDÃO, I. V. et al. Classification and predictive analysis of educational data to improve the quality of distance learning courses. In: IEEE. **2019 Workshop on Communication Networks and Power Systems (WCNPS)**. Brasília, DF, 2019. p. 1–6.

BRASIL. Lei nº 12.651, de 25 de maio de 2012. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2012. ISSN 1677-7042. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2012/Lei/L12651.htm>. Acesso em: 20 fev. 2022.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002.

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. **Neural computation**, MIT Press, v. 10, n. 7, p. 1895–1923, 1998.

DUDA, R. O.; HART, P. E. **Pattern Classification**. 2. ed. New York, USA: John Wiely and Sons, 2001.

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

FISHER, A.; RUDIN, C.; DOMINICI, F. **All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously.** 2019. 1–81 p.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001.

FRIEDMAN, J. H. Stochastic gradient boosting. **Computational statistics & data analysis**, Elsevier, v. 38, n. 4, p. 367–378, 2002.

GICIĆ, A.; SUBASI, A. Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers. **Expert Systems**, Wiley Online Library, v. 36, n. 2, p. e12363, 2019.

GRISCI, B. I.; KRAUSE, M. J.; DORN, M. Relevance aggregation for neural networks interpretability and knowledge discovery on tabular data. **Information Sciences**, Elsevier, v. 559, p. 111–129, 2021.

HAN, M. K. J.; PEI, J. **Data mining: Concepts and techniques**. 3. ed. Waltham, USA: Morgan Kaufmann, 2011.

HASTIE, T. et al. Multi-class adaboost. **Statistics and its Interface**, International Press of Boston, v. 2, n. 3, p. 349–360, 2009.

HASTIE, T. et al. **The elements of statistical learning: data mining, inference, and prediction**. Second. New York: Springer, 2009.

HAYKIN, S. **Redes neurais: princípios e prática**. Porto Alegre: Bookman Editora, 2007. 898 p.

HEBB, D. O. **The organization of behavior: a neuropsychological theory**. New York: J. Wiley; Chapman & Hall, 1949.

HOU, B.-J.; ZHOU, Z.-H. Learning with interpretable structure from gated rnn. **IEEE transactions on neural networks and learning systems**, IEEE, v. 31, n. 7, p. 2267–2279, 2020.

JONATHAN, B.; PUTRA, P. H.; RULDEVIYANI, Y. Observation imbalanced data text to predict users selling products on female daily with smote, tomek, and smote-tomek. In: IEEE. **2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)**. Bali, Indonesia, 2020. p. 81–85.

JOSE, J. T. et al. Early detection and classification of internal leakage in boom actuator of mobile hydraulic machines using svm. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 106, p. 104492, 2021.

JUNG, S. et al. Brazil's national environmental registry of rural properties: implications for livelihoods. **Ecological Economics**, Elsevier, v. 136, p. 53–61, 2017.

KALAISELVI, B.; THANGAMANI, M. An efficient pearson correlation based improved random forest classification for protein structure prediction techniques. **Measurement**, Elsevier, v. 162, p. 107885, 2020.

KARATEKIN, T. et al. Interpretable machine learning in healthcare through generalized additive model with pairwise interactions (ga2m): Predicting severe retinopathy of prematurity. In: IEEE. **2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)**. [S.l.], 2019. p. 61–66.

KARIMI, G.; HEIDARIAN, M. Facial expression recognition with polynomial legendre and partial connection mlp. **Neurocomputing**, Elsevier, v. 434, p. 33–44, 2021.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. In: ICLR. **3rd International Conference on Learning Representations, ICLR**. San Diego, USA, 2015. p. 1–15.

LI, K. et al. Predicting in-hospital mortality in icu patients with sepsis using gradient boosting decision tree. **Medicine**, Wolters Kluwer Health, v. 100, n. 19, 2021.

LICHMAN, M. et al. **Wine Dataset - UCI machine learning repository**. 2022. Disponível em: <https://archive.ics.uci.edu/ml/datasets/wine>. Acesso em: 22 fev. 2022.

LIU, D. C.; NOCEDAL, J. On the limited memory bfgs method for large scale optimization. **Mathematical Programming**, v. 45, p. 503–528, 1989.

LUNDBERG, S.; LEE, S.-I. A unified approach to interpreting model predictions. **arXiv preprint arXiv:1705.07874**, 2017.

L'ROE, J. et al. Mapping properties to monitor forests: Landholder response to a large environmental registration program in the brazilian amazon. **Land Use Policy**, Elsevier, v. 57, p. 193–203, 2016.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, n. 4, p. 115–133, 1943.

MENARDI, G.; TORELLI, N. Training and assessing classification rules with imbalanced data. **Data mining and knowledge discovery**, Springer, v. 28, n. 1, p. 92–122, 2014.

MI, J.-X.; LI, A.-D.; ZHOU, L.-F. Review study of interpretation methods for future interpretable machine learning. **IEEE Access**, IEEE, v. 8, p. 191969–191985, 2020.

MOLNAR, C. **Interpretable machine learning**. [S.l.]: Lulu. com, 2020.

PACHECO, R. et al. Will farmers seek environmental regularization in the amazon and how? insights from the rural environmental registry (car) questionnaires. **Journal of Environmental Management**, Elsevier, v. 284, p. 112010, 2021.

PIZZAIA, J. P. L. et al. Arabica coffee samples classification using a multilayer perceptron neural network. In: IEEE. **2018 13th IEEE International Conference on Industry Applications (INDUSCON)**. [S.l.], 2018. p. 80–84.

RAMCHANDANI, A.; FAN, C.; MOSTAFAVI, A. Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. **IEEE Access**, IEEE, v. 8, p. 159915–159930, 2020.

RECEITA FEDERAL. **Dados estatisticos do Cafir**. 2022. Disponível em: <https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/cadastros/portal-cnir/estatisticas-e-dados-abertos/dados-estatisticos-do-cafir>. Acesso em: 21 mar. 2022.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**. San Francisco, California, USA: Association for Computing Machinery, 2016. p. 1135–1144.

ROBLES-VELASCO, A. et al. Prediction of pipe failures in water supply networks using logistic regression and support vector classification. **Reliability Engineering & System Safety**, Elsevier, v. 196, p. 106754, 2020.

ROITMAN, I. et al. Rural environmental registry: An innovative model for land-use and environmental policies. **Land use policy**, Elsevier, v. 76, p. 95–102, 2018.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

SAGAYARAJ, M. J.; JITHESH, V.; ROSHANI, D. Comparative study between deep learning techniques and random forest approach for hrrp based radar target classification. In: IEEE. **2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)**. JCT College of Engineering and Technology Coimbatore, India, 2021. p. 385–388.

SANTOS, P. P. dos et al. Geotechnologies applied to analysis of the rural environmental cadastre. **Land Use Policy**, Elsevier, p. 105127, 2020.

Serviço Florestal Brasileiro. **Cartilha "CAR: Produzir com Respeito ao Meio Ambiente"**. 2022. Disponível em: <https://www.florestal.gov.br/component/content/article/108-publicacoes/503-cartilha-car-produzir-com-respeito-ao-meio-ambiente?Itemid=>. Acesso em: 21 mar. 2022.

Serviço Florestal Brasileiro. **Numeros do Cadastro Ambiental Rural**. 2022. Disponível em: <https://www.florestal.gov.br/numeros-do-car>. Acesso em: 21 mar. 2022.

SERVIÇO FLORESTAL BRASILEIRO. **SiCAR - Sistema Nacional de Cadastro Ambiental Rural**. 2021. Disponível em: <https://www.car.gov.br/#/baixar>. Acesso em: 25 jun. 2021.

UDDIN, J. I.; FATEMA, K.; DHAR, P. K. Depression risk prediction among tech employees in bangladesh using adaboosted decision tree. In: IEEE. **2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)**. Bhubaneswar, India, 2020. p. 135–138.

VAPNIK, V. N. **Statistical Learning Theory**. New York: Wiley, 1998.

VUČETIĆ, M.; HUDEC, M.; BOŽILOVIĆ, B. Fuzzy functional dependencies and linguistic interpretations employed in knowledge discovery tasks from relational databases. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 88, p. 15, 2020.

WANG, S.; CAO, J.; YU, P. Deep learning for spatio-temporal data mining: A survey. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, 2020.

ZHONG, H.; SONG, X.; YANG, L. Vessel classification from space-based ais data using random forest. In: IEEE. **2019 5th International Conference on Big Data and Information Analytics (BigDIA)**. Kunming, China, 2019. p. 9–12.

APÊNDICE A – Artigos Publicados em Anais de Congressos

O trabalho realizado nesta dissertação, até o presente momento, obteve duas publicações de trabalhos completos em anais de congressos de importância nacional.

Fernando Elias De Melo Borges; Danton Diego Ferreira; Antônio Carlos De Sousa Couto Júnior. Classificação de Dados do Cadastro Ambiental Rural com uso de Algoritmos de Aprendizagem de Máquina. In: **XV Simpósio Brasileiro de Automação Inteligente**, 2021, Online, 2022. v. 1.

O Cadastro Ambiental Rural (CAR) consiste em um registro público eletrônico obrigatório para todos os imóveis rurais do território brasileiro, integra informações ambientais das propriedades, auxilia o monitoramento das mesmas e no combate ao desmatamento. Entretanto, um grande número de cadastros é realizado de maneira errônea gerando dados inconsistentes, levando estes a serem cancelados e/ou a serem pedidas retificações para o devido preenchimento do cadastro. Realizar essas verificações de forma manual é deveras oneroso, uma vez que é requerida uma mão de obra especializada e o Brasil possui uma imensa quantidade de imóveis rurais. Neste contexto, este trabalho tem como objetivo fornecer um sistema inteligente baseado em aprendizagem de máquina que permita verificar e classificar os dados do CAR em aprovados ou cancelados de maneira rápida e eficaz. Para isto, três modelos de aprendizagem foram treinados utilizando dados reais de cadastros. Além da classificação, foi utilizada a ferramenta SMOTE para tratamento do desbalanceamento entre as classes. Foram gerados resultados utilizando medidas de desempenho de classificadores e realizados, também, estudos comparativos entre os métodos. Os resultados apresentados mostraram potencial uso do método em futuras predições automatizadas, atingindo índices de desempenho acima de 0.90 (90%).

Borges, Fernando Elias Melo; Ferreira, Danton Diego; Couto Júnior, Antônio Carlos De Sousa. Classificação e Interpretação de dados do Cadastro Ambiental Rural utilizando técnicas de Aprendizagem de Máquina. In: **XV Congresso Brasileiro de Inteligência Computacional**, 2021, Joinville. Anais do 15. Congresso Brasileiro de Inteligência Computacional, 2021. p. 1-7.

The Rural Environmental Registry (CAR) consists of a mandatory public electronic registry for all rural properties in the Brazilian territory, integrates environmental information of the properties, assists the monitoring of them and the fight against deforestation. However, a large number of registrations are carried out erroneously generating inconsistent data, leading these to be canceled and/or to be requested to correct the registration. Performing automatic verification of these records is important to improve the processing of records. This paper proposes an automatic classification method to approve or cancel the CAR registers with interpretation of the classifications performed. For this, four machine learning-based classifiers were tested and the results were evaluated. The model with the best performance was used to interpret the classification using the Local Interpretable Model-agnostic Explanations (LIME) algorithm. The results showed the potential of the method in future real applications.