UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE FILOSOFIA E CIÊNCIAS HUMANAS
DEPARTAMENTO DE ARQUEOLOGIA
PÓS-GRADUAÇÃO EM ARQUEOLOGIA

LUCAS BONALD PEDROSA DE SOUZA

DISTRIBUIÇÃO ESPACIAL DOS SÍTIOS ARQUEOLÓGICOS COM GRAFISMOS RUPESTRES NA BACIA HIDROGRÁFICA DO PAJEÚ, PE: UM ESTUDO DE CASO PARA APLICAÇÃO DE MODELOS PREDITIVOS BASEADOS EM ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

Recife
2023

LUCAS BONALD PEDROSA DE SOUZA

DISTRIBUIÇÃO ESPACIAL DOS SÍTIOS ARQUEOLÓGICOS COM GRAFISMOS RUPESTRES NA BACIA HIDROGRÁFICA DO PAJEÚ, PE: UM ESTUDO DE CASO PARA APLICAÇÃO DE MODELOS PREDITIVOS BASEADOS EM ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

Tese apresentada ao Programa de Pós-Graduação em Arqueologia da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Arqueologia.

Área de Concentração: Arqueologia e Conservação do Patrimônio Cultural

Orientador: Prof. Dr. Demétrio da Silva Mützenberg

Coorientador: Prof. Dr. Eduardo Krempser

Recife

2023

Catalogação na Fonte
Bibliotecário: Rodriggo Leopoldino Cavalcanti I, CRB4-1855

S729d Souza, Lucas Bonald Pedrosa de.
Distribuição espacial dos sítios arqueológicos com grafismos rupestres na bacia hidrográfica do Pajeú, PE : um estudo de caso para aplicação de modelos preditivos baseados em algoritmos de aprendizagem de máquina / Lucas Bonald Pedrosa de Souza. – 2023.
231 f. : il. ; tab. ; 30 cm.

Orientador : Demétrio da Silva Mützenberg.
Coorientador : Eduardo Krempser.
Tese (doutorado) - Universidade Federal de Pernambuco, CFCH. Programa de Pós-Graduação em Arqueologia, Recife, 2023.

Inclui referências e apêndices.

1. Arqueologia. 2. Análises espaciais. 3. MPA. 4. Aprendizagem de máquina. 5. Grafismos rupestres. I. Mützenberg, Demétrio da Silva (Orientador). II. Krempser, Eduardo (Coorientador). III. Título.

930.1 CDD (22.ed.) UFPE (BCFCH2024-048)

LUCAS BONALD PEDROSA DE SOUZA

DISTRIBUIÇÃO ESPACIAL DOS SÍTIOS ARQUEOLÓGICOS COM GRAFISMOS RUPESTRES NA BACIA HIDROGRÁFICA DO PAJEÚ, PE: UM ESTUDO DE CASO PARA APLICAÇÃO DE MODELOS PREDITIVOS BASEADOS EM ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

Tese apresentada ao Programa de Pós-Graduação em Arqueologia do Centro de Filosofia e Ciências Humanas da Universidade Federal de Pernambuco, na área de concentração Arqueologia e conservação do Patrimônio Cultural para obtenção do título de Doutor em Arqueologia.

Aprovada em: 27/02/2023

Banca examinadora:

_____

Demétrio da Silva Mutzenberg

UFPE

_____

Alencar de Miranda Amaral

UNIVASF

_____

Francisco de Assis Soares de Matos

UFPE

_____

Grégoire André Henri Marie Ghislain van Havre

UFPI

_____

Leonardo Goliatt da Fonseca

UFJF

Dedico esse trabalho a todos os profissionais e cientistas de saúde , que arriscaram suas vidas e usaram suas mentes para desenvolver uma vacina em tempo recorde para combater a Covid-19.

Dedico esta pesquisa também a meu avô, que infelizmente faleceu antes da minha defesa, mas em vida foi um árduo defensor da cultura popular e da história do povo nordestino e certamente ficaria satisfeito com meu trabalho. Muito obrigado vovô!!

AGRADECIMENTOS

Agradeço a CAPES e a FACEPE que me forneceram o auxílio financeiro necessário para a realização deste trabalho.

Agradeço a todos que fazem o Departamento de Arqueologia da Vrije Universiteit Amsterdam por terem me acolhido como um membro e me auxiliado com todo o suporte durante minha estada de 6 meses para realização do meu doutorado sanduíche.

Agradeço em especial ao Professor Philip Verhagen que me auxiliou não só na pesquisa com seus ensinamentos valiosos, mas também por ter me proporcionado ter uma visão ainda mais abrangente sobre a Arqueologia na Holanda, onde tive a oportunidade de auxiliá-lo em uma disciplina e ainda conhecer como se dão as atividades de campo.

Agradeço ao professor Grégoire Van Havre, pois sem ele o contato feito para minha ida a Holanda seria muito mais difícil. Além disso ele foi fundamental em ampliar meus horizontes com relação a esta tese, após ter tido a oportunidade de cursar sua disciplina de análises espaciais ainda na pandemia.

Agradeço ao amigo, orientador e Professor Dr. Demétrio Mützenberg pelo companheirismo e ensinamento ao longo, não só desses anos de tese, mas desde a minha graduação, me fazendo criar o gosto pela pesquisa e amar cada vez essa profissão que escolhi para a vida.

Agradeço também ao meu coorientador, Professor Eduardo Krempser, que surgiu no meio deste doutorado e teve muita paciência com minhas dúvidas primárias acerca de programação e IA e foi fundamental para a construção desta tese.

Agradeço a Luciane, pois sempre estava lá para resolver qualquer problema acadêmico.

Agradeço a Biblioteca Prof. Dra. Niède Guidon, pois nos forneceu todo o arcabouço teórico referente as nossas disciplinas e em especial, a Nelson, pois foi sempre um conselheiro e amigo de todos.

Agradeço a essas pessoas que foram fundamentais durante meus trabalhos de campo no Pajeú, em especial Seu Acioly de Afogados da Ingazeira, Dona Maria, da Serra do Pereiro, Dona Socorro, João Pedro, Pablo, Seu Pedro e Seu Beto. Sem eles as minhas atividades de campo não existiriam

Agradeço aos meus companheiros de turma de doutorado por toda contribuição nas discussões em sala de aula.

Ainda referente a turma de doutorado, vale um agradecimento especial para meus amigos Nicó e Danny, por todo o companheirismo, suporte, dicas, conselhos e etc.

Agradecimento especial também para todos os membros do GEAGT que foram fundamentais também na construção teórica e metodológica desta tese, com um agradecimento especial a Rafael, que me acompanharam durante os campos.

Agradeço a meus amigos de infância por terem me dado apoio nos momentos mais difíceis, em especial a Breno, Ganso, Heitor, Luiz e TT.

Agradeço muito a meus amigos cientistas Rogério e Bigode pela troca de conhecimentos ao longo desses anos de doutorado. Espero que nossos projetos conjuntos sigam. Agente tá enrolando bastante!!!

Agradeço à galera do Vôlei, em especial a Marcelo e Eduardo que tiveram a ideia de juntar pessoas para a prática de uma atividade física. Vocês fizeram das minhas quartas feiras mais agitadas alegres.

Agradeço a todos que fazem parte do Peregrinos, em especial a David, que me convidou para participar desta ação social que faz das minhas segundas feiras algo recompensador em poder ajudar àqueles que mais necessitam.

Agradeço a Giselle, Rayssa e Mirella pelos momentos vividos ao longo desse período. Ainda tô devendo um churrasco desde os metrado. Um dia eu pago.

Agradeço a meus amigos de turma de graduação (Carol, Duda, Endi, Ju, Newce, Kimi, Mattheus, Samba e Ramon) que podem não ter seguido no doutorado comigo, mas me apoiaram bastante ao longo desses anos.

Agradeço imensamente também aos amigos que tive a oportunidade de fazer durante meu período sanduíche, em especial a Ana, Ana Carolina, Brenno, Mari, PC, Pedro, Raquel e Rodrigo. Sem eles com certeza minha adaptação e meus dias seriam completamente diferentes.

Agradeço aos meus compadres doutores Marllon e Jéssica e a meu afilhado Lucas pelos momentos de descontração, mesmo que poucos, que tivemos a oportunidade de desfrutar.

Agradeço a Rayane, essa pessoa mais que especial que vem deixando os meus dias cada vez mais alegres e me ajudou neste final de processo de escrita da tese.

Agradeço a meus pais, Katiane e Manoel por todo apoio dado ao longo desses quatro anos, inclusive indo à campo comigo.

Agradeço a meus avós, aos que aqui estão e aos que não tiveram a oportunidade de me ver atingindo este nível acadêmico, mas que também sempre me incentivaram a trilhar o meu caminho.

“O sertanejo, é antes de tudo, um forte.” (DA CUNHA, 2013, p. 115)

RESUMO

As análises espaciais têm um destaque nos trabalhos arqueológicos, já que são fundamentais na compreensão dos contextos e na delimitação de áreas de estudos em Arqueologia. Os sítios de grafismos rupestres vêm sendo estudados no contexto espacial, desde uma escala mais abrangente até o nível específico, fazendo com que novas hipóteses possam ser traçadas a respeitos dos grupos pré-históricos que as produziram. Ultimamente, no Nordeste brasileiro, os trabalhos têm buscado observar alguns aspectos específicos das pinturas e gravuras rupestres estudadas, de modo a tentar melhor entender a dinâmica populacional do semiárido pretérito. Esta tese terá o intuito de observar esses padrões, porém a partir de uma escala macro, buscando elementos ambientais e sociais. Para realizar tal análise, serão aqui criados os chamados modelos preditivos arqueológicos (MPA), com o intuito não só de buscar os padrões, mas também, a partir deles, encontrar novos sítios arqueológicos, usando para isso uma base de dados com localidades registradas dentro da Bacia Hidrográfica do Pajeú, em Pernambuco. A modelagem aqui proposta possui um diferencial técnico, pois serão executados modelos computacionais os quais usarão de algoritmos de aprendizagem de máquina (AM), com o intuito de fornecer respostas mais graduadas e menos enviesadas aos pesquisadores. Foram criados diversos modelos, divididos em fase pré-campo, em que se usou um produto inicial para montar as estratégias para um campo, o qual teve o intuito de testar o modelo na prática e coletar mais informações. Após isso novos modelos foram criados, já com os novos dados coletados e com a inserção de novas variáveis. Os dois modelos tiveram resultados satisfatórios quando avaliadas as métricas estatísticas e ao nos debruçarmos sobre os mapas, podemos ver áreas de alta favorabilidade aparecendo como destaque, tais como a Serra do Umã, do Arapuá, Vermelha, Negra e da Baixa Verde. Para além disso, observou-se que a orientação das encostas, somada com a proximidade de áreas de grande densidade de redes de caminhos ótimos que usaram a declividade e a insolação como superfície de custo, tiveram uma importância significativa nos resultados do modelo. Por fim fica claro que o uso dos MPA contribuiu para um melhor entendimento acerca dos padrões associados aos sítios arqueológicos desta natureza, porém para se ter melhores resultados, mais informações precisam ser coletadas e inseridas a base.

Palavras-chave: análises espaciais; MPA; aprendizagem de máquina; grafismos rupestres.

ABSTRACT

Spatial analyzes are highlighted in archaeological work, as they are fundamental in understanding contexts and delimiting areas of study in Archaeology. Rock art sites have been studied in a spatial context, from a broader scale to a specific level, allowing new hypotheses to be drawn about the prehistoric groups that produced them. Lately, in the Brazilian Northeast, works have sought to observe stylistic and technical patterns in the paintings and engravings studied, trying to better understand the past population dynamics of brazilian semi-arid. This thesis will aim to observe these patterns, from a macro scale, looking for environmental and social elements. To carry out such analysis, the so-called archaeological predictive models (APM) will be created, with the aim not only to looking for patterns, but also, to find new archaeological sites, using a database with locations registered within the Pajeú Watershed, in Pernambuco. The modeling proposed here has a technical differential. The computational models will be executed using machine learning (ML) algorithms, in order to provide more graduated and less biased responses to researchers. Several models were created, divided into a pre-field phase, in which an initial product was used to assemble strategies for a field, which aimed to test the model in practice and collect more information. After that, new models were created, already with the new data collected and with the insertion of new features. The two models had satisfactory results when the statistical metrics were evaluated and when looking at the maps, we can see areas of high probability appearing as highlights, such as Serra do Umã, do Arapuá, Vermelha, Negra and Baixa Verde. In addition, it was observed that the orientation of the slopes, together with the proximity of areas of great density of networks of optimal paths that used the slope and the insolation as a cost surface, had a significant importance in the results of the model. Finally, it is clear that the use of APM contributed to a better understanding of the patterns associated with archaeological sites of this nature, but in order to have better results, more information needs to be collected and entered into the database.

Key words: spatial analyses; APM; machine learning; rock art

RESUMÉE

Les analyses spatiales sont mises en avant dans les travaux archéologiques, car elles sont fondamentales pour comprendre les contextes et délimiter les domaines d'étude en Archéologie. Les sites d'art rupestre ont été étudiés dans un contexte spatial, d'une échelle plus large à une échelle spécifique, permettant de formuler de nouvelles hypothèses sur les groupes préhistoriques qui les ont produits. Dernièrement, dans le Nord-Est brésilien, des travaux ont cherché à observer les schémas stylistiques et techniques des peintures et gravures rupestres étudiées, afin de essayer de mieux comprendre la dynamique des populations de la région semi-aride passée. Cette thèse visera à observer ces schémas, mais à une échelle macro, en recherchant des éléments environnementaux et sociaux. Pour effectuer une telle analyse, les soi-disant modèles prédictifs archéologiques (MPA) seront créés ici, dans le but non seulement de rechercher des modèles, mais aussi, à partir d'eux, de trouver de nouveaux sites archéologiques, en utilisant une base de données avec des emplacements enregistrés dans le bassin hydrographique de Pajeú, à Pernambuco. La modélisation proposée ici a un différentiel technique, car des modèles informatiques seront exécutés qui utiliseront des algorithmes d'apprentissage automatique (AA), afin de fournir des réponses plus graduées et moins biaisées aux chercheurs. Plusieurs modèles ont été créés, divisés en une phase de pré-terrain, dans laquelle un produit initial a été utilisé pour assembler des stratégies pour une activitée de terrain, qui visait à tester le modèle dans la pratique et à collecter plus d'informations. Après cela, de nouveaux modèles ont été créés, déjà avec les nouvelles données collectées et avec l'insertion de nouvelles variables. Les deux modèles ont donné des résultats satisfaisants lors de l'évaluation des mesures statistiques et en regardant les cartes, nous pouvons voir des zones de haute favorabilité apparaître comme des points forts, comme Serra do Umã, do Arapuá, Vermelha, Negra et Baixa Verde. De plus, il a été observé que l'orientation des pentes, ainsi que la proximité des zones de grande densité de réseaux de chemins optimaux qui utilisaient la pente et l'insolation comme surface de coût, avaient une importance significative dans les résultats du modèle. Enfin, il est clair que l'utilisation de les MPA a contribué à une meilleure compréhension des schémas associés aux sites archéologiques de cette nature, mais pour obtenir de meilleurs résultats, davantage d'informations doivent être collectées et saisies dans la base de données.

Mots clés :analyses spatiales, MPA, apprentissage automatique, art rupestre

LISTA DE FIGURAS

Figura 1: Breve linha do tempo dos estudos com MPA.....35

Figura 2: Localização dos territórios indígenas do Pajeú.....61

Figura 3: Mapa de distribuição dos grupos indígenas do Estado de Pernambuco no século XVIII.....64

Figura 4: Mapa de localização das missões jesuíticas do submédio São Francisco.....67

Figura 5: Mapa Hipsométrico de Localização da Bacia Hidrográfica do Pajeú e seus sítios arqueológicos.....69

Figura 6: Painel com Gravuras Rupestres do sítio Pedra do Riacho da Velha, no município de Flores.....71

Figura 7: Gravura rupestre do sítio arqueológico Mãe D'água I, no município de Floresta.. .72

Figura 8: Foto de grafismos puros pintados no sítio Pedra do Catolé, no município de Carnaubeira da Penha.....73

Figura 9: Marcas de mão pintadas visualizadas no sítio Pedra da Cuia, no município de Carnaubeira da Penha.....74

Figura 10: Precipitações médias e temperaturas médias do Pajeú. A) Volume precipitação anual. B) Temperatura média anual.....76

Figura 11: Litoestratigrafia do Pajeú.....78

Figura 12:Seção estratigráfica Lagoa Uri de Cima.....81

Figura 13: Tipo de movimentos na construção de uma rede de caminhos de menor custo.. 94

Figura 14: Mapa de caminhos ótimos baseados na declividade como superfície de custo.. 95

Figura 15: Modelo de regressão linear.....103

Figura 16: gráficos de linhas de ajuste do modelo de regressão linear (à esquerda) e percepção dos erros associados a linha melhor ajustada (à direita).....104

Figura 17: Função logística ou sigmoide.....105

Figura 18: Representação de uma função quadrática com o ponto de erro mínimo.....107

Figura 19: Árvore de decisão Conjunto de Dados Isis.....110

Figura 20: Rede Neuronal Perceptron Multicamadas com uma camada escondida.....113

Figura 21 - Esquema representando o processo de leave-one-out.....116

Figura 22: Exemplo de matriz de confusão.....117

Figura 23: Representação gráfica dos valores SHAP.....120

Figura 24: Gráficos com a distribuição dos resultados referente as avaliações de cada uma das sementes aleatórias para o modelo de RNA.....124

Figura 25: MPA Inicial com o método de reamostragem ADASYN (Regressão Logísitica) 126

Figura 26: MPA Inicial com o método de reamostragem Kmeans + SMOTE (Regressão Logísitica).....127

Figura 27: MPA Inicial com o método de reamostragem SMOTE (Regressão Logísitica)....128

Figura 28: MPA Inicial com o método de reamostragem ADASYN (Floresta Aleatória).....131

Figura 29: MPA Inicial com o método de reamostragem Kmeans + SMOTE (Floresta Aleatória).....132

Figura 30: MPA Inicial com o método de reamostragem SMOTE (Floresta Aleatória).....133

Figura 31: MPA Inicial (Floresta Aleatória Balanceada).....136

Figura 32: MPA Inicial (Bagging Balanceado).....137

Figura 33: MPA Inicial com o método de reamostragem ADASYN (Ensemble).....140

Figura 34: MPA Inicial com o método de reamostragem Kmeans + SMOTE (Ensemble)....141

Figura 35: MPA Inicial com o método de reamostragem SMOTE (Ensemble).....142

Figura 36: Histograma de importâncias das variáveis referentes ao modelo de floresta aleatória.....144

Figura 37: Histograma de importâncias das variáveis referentes ao modelo de FAB.....145

Figura 38: Histograma de importâncias das variáveis referentes ao modelo de regressão logística e *smote* como método de reamostragem.....145

Figura 39: Áreas prospectadas durante as atividades de campo de junho de 2021.....147

Figura 40: Sítio não cadastrado com grafismos rupestres na Serra das Flores, Carnaubeira da Penha.....148

Figura 41: Gravuras nas margens do Rio Pajeú.....150

Figura 42: Mapa de densidade dos caminhos ótimos do Pajeú usando a declividade como superfície de custo.....152

Figura 43: Mapa de Proximidade dos caminhos ótimos do Pajeú usando a insolação como superfície de custo.....153

Figura 44: MPA_Pajeu_1.0 com o método de reamostragem ADASYN (regressão logística).....154

Figura 45: MPA_Pajeu_1.0 com o método de reamostragem Kmeans + SMOTE (regressão logística).....155

Figura 46: MPA_Pajeu_1.0 com o método de reamostragem SMOTE (regressão logística).....156

Figura 47: MPA_Pajeu_1.0 com o método de reamostragem ADASYN (floresta aleatória) 158

Figura 48: MPA_Pajeu_1.0 com o método de reamostragem Kmeans + SMOTE (floresta aleatória).....159

Figura 49: MPA_Pajeu_1.0 com o método de reamostragem SMOTE (floresta ale-atória) 160

Figura 50: MPA_Pajeu_1.0 com o método de reamostragem ADASYN (RNA Perceptron) 162

Figura 51: MPA_Pajeu_1.0 com o método de reamostragem Kmeans SMOTE (RNA Perceptron).....163

Figura 52: MPA_Pajeu_1.0 com o método de reamostragem SMOTE (RNA Perceptron).....164

Figura 53: MPA_Pajeu_1.0 (*Bagging* balanceado).....166

Figura 54: MPA_Pajeu_1.0 (Floresta Aleatória Balanceada).....167

Figura 55: MPA_Pajeu_1.0 com o método de reamostragem ADASYN (Ensemble).....169

Figura 56: MPA_Pajeu_1.0 com o método de reamostragem Kmeans + SMOTE (Ensemble).....170

Figura 57: MPA_Pajeu_1.0 com o método de reamostragem SMOTE (Ensemble).....171

Figura 58: Gráfico de valores *SHAP* de importância das variáveis (Regressão Logística).....174

Figura 59: Gráfico de valores *SHAP* de importância das variáveis (RNA Perceptron).....175

Figura 60: Gráfico de valores *SHAP* de importância das variáveis (Floresta Aleatória).....175

Figura 61: Gráfico de valores *SHAP* de importância das variáveis (FAB).....176

Figura 62: Localização do PARNA do Catimbau e locação dos pontos de origem /destino.....179

Figura 63: MPA PARNA do Catimbau com o método de reamostragem Adasyn (Ensemble).....180

Figura 64: MPA PARNA do Catimbau com o método de reamostragem Kmeans + Smote (Ensemble).....181

Figura 65: MPA PARNA do Catimbau com o método de reamostragem Smote (Ensemble).....182

Figura 66: Serra do Pereiro.....184

Figura 67: Sítio Pedra do Riacho da Velha.....185

Figura 68: (em cima) Imagem geral do suporte com pinturas no sítio Serrote Verde. (abaixo) Imagem tratada no aplicativo Rock Art Enhancer, onde é possível ver claramente os antropomorfos dispostos no painel.....186

Figura 69: à esquerda: Sítio Lajedo dos Porfírios, ainda não cadastrado; à direita sítio Pedra Gravada, já presente na base de dados.....187

Figura 70: Abrigo na Serra da Tiúba.....188

Figura 71: à esquerda: imagem de um dos grafismos encontrados no sítio da Serra das Flores. À direita: foto editada no software Rock Art Enhancer.....189

Figura 72: Mapa de localização dos sítios encontrados durante a prospecção realizada no ano de 2022.....190

Figura 73: MPA treinado exclusivamente com os sítios de gravuras usando o método de reamostragem Kmeans+ SMOTE (Regressão logísitca).....192

Figura 74: MPA treinado exclusivamente com os sítios de gravuras usando o método de reamostragem Kmeans+ SMOTE (Regressão logísitca).....193

Figura 75: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (Regressão logística).....195

Figura 76: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (RNA – Perceptron MC).....196

Figura 77: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (Floresta Aleatória).....197

Figura 78: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (FAB).....197

Figura 79: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (Bagging Balanceado).....198

Figura 80: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (Ensemble).....199

Figura 81: Gráfico de beeswarm contendo a importância das variáveis nos modelos Regressão logística (A); RNA Perceptron (B); Floresta Aleatória (C); FAB (D).....201

LISTA DE TABELAS

Tabela 1: Métricas estatísticas de avaliação para o modelo de regressão logística.....129

Tabela 2: Matriz de confusão regressão lógistica.....130

Tabela 3: Métricas estatísticas de avaliação para o modelo de floresta aleatória.....134

Tabela 4: Matriz de confusão floresta aleatória.....135

Tabela 5: Métricas estatísticas de avaliação para os modelos de floresta aleatória balanceada e *bagging* balanceado.....138

Tabela 6:Matriz de confusão FAB e *Bagging* Balanceado.....139

Tabela 7: Métricas estatísticas de avaliação para o modelo ensemble.....143

Tabela 8: Matriz de confusão modelo Ensemble.....143

Tabela 9: Métricas estatísticas de avaliação para o modelo final de regressão logísitca.....157

Tabela 10: Métricas estatísticas de avaliação para o modelo final de floresta aleatória.....161

Tabela 11: Métricas estatísticas de avaliação para o modelo final de RNA Perceptron multicamadas.....165

Tabela 12: Métricas estatísticas de avaliação para os modelos finais de *bagging* balanceado e FAB.....168

Tabela 13: Métricas estatísticas de avaliação para o modelo ensemble.....172

Tabela 14: matriz de confusão dos modelos finais.....173

Tabela 15: Métricas estatísticas de avaliação para os modelos do MPA_Pajeu_1.1.....203

Tabela 16: Hiperparâmetros usados para os cálculos dos modelos.....227

Tabela 17: Quadro das variáveis usadas no modelo.....228

LISTA DE ABREVIATURAS E SIGLAS

AM	**Aprendizagem de Máquina**
AP	**Antes do Presente**
CHESF	**Companhia Hidrelétrica do São Francisco**
CNSA	**Cadastro Nacional de Sítio Arqueológicos**
CODEVASF	**Companhia de Desenvolvimento dos Vales dos Rios São Francisco e Parnaíba**
CPRM	**Companhia de Pesquisa de Recursos Mineirais**
CRM	**Cultural Resource Management**
IA	**Inteligência Artificial**
IBM	**International Business Machines Corporation**
IPHAN	**Instituto do Patrimônio Histórico e Artístico Nacional**
IPT	**Índice de Posição Topográfica**
MLP	**Multilayer Perceptron**
MPA	**Modelo Preditivo em Arqueologia**
PARNA	**Parque Nacional**
SIG	**Sistema de Informação Geográfica**
SRTM	**Shuttle Radar Topography Mission**
SUDENE	**Superintendência do Desenvolvimento do Nordeste**
ZCIT	**Zona de Convergência Intertropical**

SUMÁRIO

1 INTRODUÇÃO 24

2 APORTES TEÓRICOS 32

2.1 Analisando o espaço arqueológico através do mundo digital 32

2.1.1 Os Modelos Preditivos 34

2.1.2 Os Modelos Preditivos em Arqueologia (MPA): uma linha do tempo 35

2.1.3 Alertas no uso da modelagem preditiva 37

2.2 A Inteligência Artificial 43

2.1.1 A Arqueologia e a IA 46

2.1.2 Aplicações na Arqueologia 46

2.3 O Sítio Arqueológico 47

2.4 Grafismos Rupestres 51

2.4.1 O caráter comunicativo dos grafismos rupestres 54

2.4.2 Técnicas de análise e as tradições rupestres 55

3 A BACIA DO PAJEÚ E SEUS CONTEXTOS 59

3.1 O Contexto etno-histórico 59

3.2 Contexto arqueológico no Pajeú 65

3.2.1 Os sítios rupestres da região 70

3.3 Contexto ambiental 75

3.4 Contexto paleoambiental 79

4 MÉTODOS 83

4.1 Um pouco mais sobre o SIG na Arqueologia 84

4.2 Sensoriamento Remoto 85

4.3 A coleta de dados e a delimitação da área de estudo 87

4.4 Levantando as variáveis da pesquisa 89

4.4.1 Inclinação 90

4.4.2 Aspecto ou Orientação das encostas 90

4.4.3 Insolação 91

4.4.4 Índice de Posição Topográfica (IPT) 91

4.4.5 Perfil de Curvatura 91

4.4.6. Distância para os principais cursos d'água 91

4.4.7 Distância para as litologias 92

4.4.8 Índice de Visibilidade (IV) 92

4.4.9 Acessibilidade 93

4.5 Discutindo as variáveis da pesquisa 95

4.6 Calculando as estatísticas das variáveis 96

4.7 A construção do MPA baseado em algoritmos de AM 97

4.7.1 Exportando os dados para o Python 98

4.7.2 Estruturando a AM 98

4.7.3 Divisão dos conjuntos de dados 99

4.7.4 Conjunto de dados desbalanceados 100

4.8 Os algoritmos 101

4.8.1 Regressão linear 102

4.8.2 Regressão logística 104

4.8.3 Árvores de decisão e florestas aleatórias 108

4.8.4 As redes neuronais artificiais (RNA) 112

4.8.5 Outros algoritmos 114

4.9 Avaliando o modelo 114

4.9.1 Avaliando as variáveis do modelo 118
4.9.2 Interpretando as variáveis 121
4.10 Testando o modelo em campo 121
5. MPA_PAJEÚ_0.1: O PRIMEIRO MODELO 123
5.1 Os primeiros resultados 125
. 5.2 A importância das variáveis nos modelos iniciais 143
. 5.3 O teste em campo 146
6. MPA_PAJEU_1.0: MELHORANDO OS MODELOS 151
6.1 O desenvolvimento dos modelos 151
6.1.2 Os modelos gerados 153
6.2 A importância das variáveis no MPA_Pajeu_1.0 174
7. PONDO OS MODELOS À PROVA 178
7.1 Calculando as estatísticas do modelo 179
7.2 Discutindo os resultados 183
7.3 A segunda visita a campo 183
7.4 Discutindo os resultados do campo 190
7.5 Separando as gravuras das pinturas 191
7.6 MPA_Pajeu_1.1: Os modelos pós-defesa 194
7.6.1 A importância das variáveis 200
7.6.2 Avaliando estatisticamente o MPA_Pajeu_1.1 202
8 CONSIDERAÇÕES FINAIS 204
REFERÊNCIAS 207

APÊNDICE A - Códigos e dados..... 226

APÊNDICE B – hiperparâmetros do modelo..... 227

APÊNDICE C – variáveis do modelo..... 228

APÊNDICE D – gráficos de sementes do modelo..... 230

1 INTRODUÇÃO

Nem só do passado vive a Arqueologia, e este trabalho tem o intuito de testar novas metodologias de análise espacial, criando modelos preditivos arqueológicos baseados em algoritmos de aprendizagem de máquina, de modo a encontrar novos sítios arqueológicos na região da Bacia do Pajeú, PE, buscando variáveis de cunho social e ambiental que mais se relacionam com os sítios de grafismos rupestres lá encontrados.

Os estudos espaciais em Arqueologia visam entender como se deram as relações entre os grupos humanos do passado e o ambiente em que estes viviam. Nos primórdios da ciência arqueológica, quando as pesquisas ainda estavam ligadas a ideia do colecionismo, os elementos de disciplinas como a Cartografia foram os primeiros a serem usados para encontrar e melhor compreender a respeitos de sítios arqueológicos trabalhados..

Com o passar dos anos e os avanços tecnológicos, os trabalhos em campo e as análises laboratoriais, passaram a gerar uma quantidade de dados muito grande. . No que diz respeito às análises espaciais especificamente, houve melhorias na topografia, no sensoriamento remoto, na fotogrametria e na geofísica, o que contribuiu para se ter acesso a uma gama de informações que não eram acessíveis de início .

É a partir disto que vão surgindo novas formas de analisar o registro arqueológico e o ambiente que o circundava. Uma delas, os modelos preditivos arqueológicos (MPA), muito popularizados nos anos 60, quando da realização de alguns trabalhos nos EUA, incorporados à *Cultural Resource Management* (CRM), equivalente a Arqueologia de Contrato no Brasil. Como este tipo de pesquisa demanda uma celeridade maior, pois lida diretamente com a Engenharia civil e obras de impacto infraestrutural, a necessidade por metodologias que pudessem fornecer dados de forma mais rápida e ao mesmo tempo com qualidade, fez dos MPA, uma grande ferramenta metodológica .

Ainda no final do século XX, com o avanço da computação, tanto em seus componentes físicos como nos *softwares*, fez com que a modelagem preditiva ganhasse alguns aditivos. O principal deles é o uso da Inteligência Artificial (IA) em aplicações de sistemas baseados em conhecimento, que na Arqueologia estão relacionados a avaliações de hipóteses,

simulações e predições de sítios arqueológicos e classificação de vestígios, como foi o caso do pioneiro programa criado por Doran para classificar e seriar os achados do cemitério de La Tène, na Suíça . Um dos ramos da IA que vem sendo muito utilizado ultimamente, é o chamado Aprendizagem de Máquina (AM), que tem no seu escopo uma série de algoritmos capazes de "aprender" com uma determinada camada de dados e assim aplicar seu conhecimento em informações ainda desconhecidas .

Um dos trabalhos de aplicação dos MPA, que usa da IA nos dias atuais, é o *Minnesota Model* ou *MnModel*[1]. Este vem sendo aprimorado e usado pelo Governo do Estado de Minnesota, nos EUA, há pelo menos 20 anos. Em 7 anos o projeto custou 4 milhões de dólares aos cofres públicos, porém, por ano, ele conseguiu economizar cerca de 3 milhões , o que mostra o quão vantajoso é investir em projetos deste tipo.

Esta pesquisa, especificamente, trabalhará com os MPA atrelados aos algoritmos de AM, e além disso, partirá de uma dicotomia ausência/presença, ou seja, teremos como dados iniciais informações referentes aos sítios arqueológicos e também relativas aos locais sem vestígio rupestre algum e a partir disto construiremos nossa base metodológica.

Esta dicotomia causa um certo problema, já que trabalharemos com muito mais locais de ausência, do que o contrário, o que faz a base de dados deste trabalho ser definida como "desbalanceada" . Esta pesquisa, portanto, buscará também soluções para trabalhar com este tipo de problema metodológico.

Esta pesquisa não se resume aos aspectos metodológicos, afinal para aplicação de tais técnicas, é preciso não só dos dados arqueológicos, mas também de todo o contexto ambiental em que os registros rupestres estão inseridos. A principal área de estudo aqui observada, é a região da Bacia Hidrográfica do Pajeú, na porção central do Estado de Pernambuco.

O Sertão do Pajeú, como popularmente é conhecida a região da maior Bacia Hidrográfica do Estado de Pernambuco, ficou eternizada na letra de Luiz Gonzaga e Zé Dantas,

[1] Mais informações sobre o referido modelo podem ser encontradas no link a seguir: https://www.dot.state.mn.us/mnmodel/archaeology.html

que relata os cursos de suas drenagens na música “Riacho do Navio” . Além disso, no início do século XX, uma rede de alianças fortes, criada com o intuito de se proteger de seus desafetos, foi lá construída, por Lampião e seu bando de cangaceiros .

A região apesar de localizada em uma porção de clima semiárido, com chuvas irregulares, guarda em seu território, muitos vestígios de histórias que antecederam os dois personagens acima citados. Segundo alguns trabalhos arqueológicos realizados na região , as ocupações nas áreas remontam há milhares de anos, com presenças de cerâmica e materiais líticos. Além disso, estudos baseados em relatos etnográficos e cartografia histórica , atestam que a região do Pajeú já era lugar de uma diversidade étnica grande, quando da chegada dos europeus.

Espacialmente falando, a região de estudo deste trabalho abrange os perímetros da Bacia Hidrográfica do Rio Pajeú. Esta, por sua vez, engloba uma área de 16.838,70 km². Está localizada na porção central do estado, entre as coordenadas 07º 16’ 20” e 08º 56’ 01” de latitude sul e 36º 59’ 00” e 38º 57’ 45” de longitude oeste.

Esta pesquisa terá como principal objeto de estudo, os sítios com grafismos rupestres encontrados nos limites da Bacia, registrados desde o ano de 2014, com o início do Projeto de caracterização e documentação dos sítios rupestres no semiárido de Pernambuco. Segundo Pessis et al , este projeto foi criado com o intuito de catalogar o máximo de sítios possíveis, sobretudo no Agreste e Sertão pernambucano, devido ao alto grau de deterioração, sofrido pelos mesmos. Estes servirão de base para a construção de uma metodologia de análise espacial que cria modelos preditivos arqueológicos, com o intuito não só de achar novos vestígios rupestres no local, mas também com interesse de melhor compreender o contexto ambiental associado aos sítios deste tipo.

O Projeto de caracterização e documentação dos sítios rupestres no semiárido de Pernambuco, catalogou 155 sítios de registro rupestre ao longo da Bacia do Rio Pajeú, dos quais 71 são de pintura, 54 de gravura e 30 que envolvem as duas técnicas. Estes estão dispostos desde locais próximos a nascente, na divisa entre os estados de PE e PB, até a foz, no encontro com o Rio São Francisco.

A região do Vale do São Francisco, que se apresenta como uma das áreas de grande destaque na ocupação e deslocamento dos primeiros grupos humanos da América, apresenta rastros da presença humana ao longo de todo o curso, desde a nascente, na Serra da Canastra, em MG, até sua foz, entre os estados de AL e SE .

O médio curso do São Francisco foi alvo de diversas pesquisas, sobretudo nos anos 80, quando da construção das hidrelétricas de Itaparica e Xingó, proporcionando o achado de diversos sítios arqueológicos. Em Itaparica, localizada nas proximidades da foz do Rio Pajeú, foram evidenciados vestígios que indicavam a presença de acampamentos temporários nos terraços fluviais e de ocupações mais duradouras, nos abrigos sob rocha às margens do Rio São Francisco e nas próprias ilhas que este acomoda .

Devido a essa presença de sítios com alta concentração de vestígios, Etchevarne , sugere que os grupos pré-históricos se aproveitavam dos canais de rios intermitentes, como é o caso do Pajeú e também do Moxotó e Ipanema , para explorar a caatinga, mas sempre retornavam às margens do São Francisco.

A região dos rios Ipanema e Moxotó citados acima, são afluentes do São Francisco, vem sendo alvo de pesquisas na área do grafismo rupestre nos últimos anos , onde tem sido observado uma recorrência nos padrões estilísticos nas pinturas rupestres da Tradição Nordeste encontradas na região.

Matos (2019), com um trabalho que englobou diferentes áreas do semiárido nordestino, observou através de uma análise de um caminho de menor custo, que usava como ponto de partida o Seridó Potiguar e indo até a área do PARNA do Catimbau e T.I Kapinawá a partir dos dados dos sítios arqueológicos com registros de pinturas da Tradição Nordeste destas regiões. No estudo foi observado que boa parte dos sítios com registros parietais estavam relacionados aos cursos d’água. Além disso, os rios Moxotó e Ipanema, tinham nas suas cabeceiras, relações diretas com as cabeceiras do Rio Paraíba, à norte, e este, por sua vez, era usado como caminho para se chegar ao Seridó e na cabeceira dos principais rios da Região.

Os sítios da tradição Nordeste analisados por Matos (2019), têm como característica suas representações picturais de cenas do cotidiano dos grupos pré-históricos, é amplamente descrita nos trabalhos arqueológicos sobre o tema . Há ocorrências desta, desde o Norte de

Minas , passando pela Serra da Capivara (PI) e em boa parte do Vale do São Francisco (MARTIN, 1998).

Com grande ocorrência no Nordeste brasileiro também, a Tradição Agreste, com seus painéis mais estáticos, desenhos isolados, marcas de mão e grafismos puros, tem como primeira área de dispersão, possivelmente a porção central de Pernambuco, sendo facilmente encontrada em sítios arqueológicos do Pajeú .

É importante salientar que estas diferenças técnico-estilísticas usadas para agrupar os registros rupestres em "tradições", não leva em consideração as questões cronológicas, afinal ainda não temos métodos diretos de datação das pinturas, porém através de datações de materiais associados, a Tradição Agreste e Nordeste, devem ter se disseminado em diferentes cronologias .

Voltando ao Rio Pajeú e sua bacia, e localizado nesse contexto de diversos achados arqueológicos, podemos depreender então, que o local tinha uma importância para os grupos pré-históricos, sobretudo ao levarmos em consideração o caráter comunicador dos grafismos, no sentindo de passar alguma informação a quem os avistasse (PESSIS, 1989; CISNEIROS 2008; MATOS 2019). Além disso, sua proximidade com Rio São Francisco e os próprios dados etno-históricos , reforçam o destaque desta Região no sentindo de ampliar o conhecimento sobre a pré-história nordestina.

A Bacia do Pajeú, configura-se, portanto, como um local de grande potencial para a compreensão das dinâmicas espaciais dos grupos pré-históricos que habitaram o semiárido nordestino, haja vista sua proximidade às redes de drenagem que compõe às cabeceiras dos rios Paraíba, Moxotó e Ipanema, regiões já atestadas como pontos de passagem no deslocamento entre o Seridó e o PARNA do Catimbau, por exemplo.

Afora isso, em uma análise prévia dos dados coletados por Pessis et al (2018), observa-se que as gravuras e pinturas rupestres aparecem com certo equilíbrio, na paisagem do Pajeú e os grafismos puros, aqueles caracterizados por formas geométricas ou abstrações, são mais recorrentes do que outros elementos.

Pensando na predominância de grafismos puros, em detrimento, das grandes cenas comuns à Tradição Nordeste, e no caráter comunicador do registro rupestre, esta pesquisa parte

da seguinte pergunta: Existem padrões ambientais e/ou sociais recorrentes relacionados às pinturas e gravuras rupestres feitas pelos grupos pré-históricos que ocuparam a Bacia Hidrográfica do Pajeú?

A Arte Rupestre como um todo, é um indicador fiável do deslocamento e dispersão de técnicas na Pré-História. A diversidade técnica e cenográfica dos grafismos rupestres, podem sugerir atividades em períodos históricos distintos. O que mostra o Pajeú como sendo uma área onde pode ter havido deslocamento de grupos, durante vários períodos da pré-história brasileira.

Com os pontos levantados acima e ciente da pergunta proposta para esta pesquisa chega-se a seguinte solução: A distribuição espacial, atrelada à diferença técnica nos tipos de grafismos encontrados, indicam que as produções dos grafismos rupestres realizadas no Pajeú, foram feitas seguindo critérios socioambientais específicos.

Além do problema arqueológico que delimita esta pesquisa, temos também um questionamento de cunho metodológico, que visa contornar algumas dificuldades encontradas devido ao conjunto de dados aqui usado. Ao trabalhar em uma dicotomia ausência/presença (sítio/não sítio), enfrentaremos um problema de desequilíbrio entre os dados, já que haverá mais áreas sem nenhum vestígio, do que o contrário. Portanto, este trabalho necessita lidar com um conjunto de dados em que uma classe é muito maior que a outra.

Esse trabalho tem como objetivo principal buscar padrões de recorrência em elementos ambientais e sociais atrelados à sítios rupestres. Para além disso, esta pesquisa procurou desenvolver de forma específica alguns pontos:

1. um modelo preditivo arqueológico com o auxílio de técnicas computacionais;
2. utilizar variáveis de cunho social nos modelos preditivos gerados, de modo a diminuir o determinismo ambiental;
3. compreender a distribuição espacial dos sítios arqueológicos com grafismos rupestres no Pajeú;
4. Testar e aprimorar técnicas para lidar com conjuntos de dados desbalanceados;
5. Comparar os dados espaciais analisados neste trabalho com o de outras pesquisas nas áreas vizinhas, de modo a ter uma visão mais ampla a respeito do contexto.

Diante do que foi exposto, em termos de relevância, este trabalho pode ser enquadrado em três pontos. Primeiramente, uma relevância científica, pois ela pode preencher lacunas sobre a pré-história da região. Ela também traz uma relevância tecnológica, pois ao final da produção do modelo preditivo, há um interesse que o mesmo possa ser usado em outras pesquisas que trabalhem num contexto similar, o que poderá impulsionar ainda mais a Arqueologia no semiárido nordestino. Por fim, ela também pode ter uma importância econômica, já que existem pesquisas de Arqueologia ligados a obras de infraestrutura, e com dados de modelos preditivos, é possível dar uma maior celeridade a estes trabalhos, e consequentemente diminuir os custos. Com a promulgação da Instrução Normativa nº1/2015 do Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN) inclusive, passou a ser demandado aos profissionais arqueólogos, que realizassem modelos preditivos que pudessem indicar locais com maiores favorabilidades de encontrar sítios arqueológicos, de modo a direcionar as prospecções em campo (BRASIL, 2015).

De forma resumida, esta tese estará organizada em mais 6 capítulos além desta introdução. Iniciando pela 2ª seção, onde serão apresentados os conceitos centrais que guiam o uso das aplicações computacionais na Arqueologia; as análises espaciais, no qual são incluídos sobretudo os modelos preditivos arqueológicos; também serão inseridos os principais conceitos da aprendizagem de máquina utilizados neste trabalho; e por fim, um levantamento dos principais conceitos por trás dos estudos dos grafismos rupestres.

No 3º capítulo da tese o foco será em entender o contexto no qual o sítio arqueológico está inserido. Três partes principais compõe a seção em questão, com início no contexto etnohistórico, seguido de um contexto arqueológico e por fim de uma caracterização ambiental e paleoambiental do Pajeú.

O 4º capítulo por sua vez trata da metodologia aplicada nesta tese, apresentando não somente um passo a passo sobre o que foi feito até se chegar no produto final, mas também elucidando mais algumas conceituações, que ficaram mais bem aplicadas ao longo da explicação dos métodos escolhidos para integrar o trabalho.

O 5º capítulo da tese tem como objetivo apresentar e discutir os resultados obtidos no modelo inicial gerado. Aqui não foram usadas todas as opções metodológicas expressas no capítulo 3, mas o produto foi utilizado como guia da primeira atividade de campo conduzida

durante o mês de junho de 2021, com o intuito de atestar a qualidade da modelagem e coletar mais informações, que posteriormente seriam incluídas no modelo final.

O 6º capítulo traz os resultados do modelo final, já com os dados coletados no campo de junho de 2021 e com a inserção de todas as variáveis e o uso de toda a metodologia relatada, através da apresentação dos dados estatísticos e dos gráficos de importância das variáveis.

O 7º capítulo relata a avaliação prática dos produtos gerados. Foram criados mapas de predição na região do PARNA do Catimbau, usando como base os dados de treino do Pajeú. Além disso aqui são apresentadas fotos dos locais visitados durante a última atividade de campo, em 2022 e os sítios que foram encontrados durante esta visita.

O 8º capítulo é um fechamento dos estudos que aqui serão apresentados e já com o olhar para o futuro, nas possibilidades que esse tipo de trabalho pode ser ampliado no âmbito do próprio Pajeú, de modo a aprimorar ainda mais as informações obtidas e também extrapolar os limites da área de estudo, tentando aplicá-lo a mais áreas dentro do semiárido nordestino.

2 APORTES TEÓRICOS

A Arqueologia é uma ciência das mais diversas quando o assunto é dialogar com outras áreas do conhecimento. Hoje vemos como o avanço de áreas como a Biologia, Geologia, Química, Física, Medicina entre outras áreas, tem contribuído para a expansão da Ciência Arqueológica .

Nesta tese a Arqueologia será tratada como esta área do conhecimento plural, e abarcaremos conceitos advindos da Geografia, Computação e Estatística, as quais serão base para a construção de toda a base de ideias, que parte desde a definição de sítio arqueológico até a aplicação de um algoritmo de AM.

2.1 Analisando o espaço arqueológico através do mundo digital

Em face desta pluralidade de conceitos, esta tese tem o interesse de encontrar aspectos do ambiente que possam estar relacionados com os sítios arqueológicos de grafismos rupestres do Pajeú, e para isso será necessário a utilização de diversas ferramentas que irão auxiliar nesta tarefa de analisar os vestígios e os contextos associados.

A Arqueologia, assim como todas as ciências na atualidade, foram impulsionadas pelas inovações tecnológicas proporcionadas pelo avanço da capacidade computacional. Um dos grandes desafios do arqueólogo sempre foi o de lidar com os dados. Nos primórdios desta ciência, estes eram poucos, porém com metodologias de campo cada vez mais robustas, as informações coletadas passam a crescer muito e demandam uma gestão mais cuidadosa, para tal, a utilização de meios digitais às pesquisas arqueológicas, torna-se imprescindível .

A união da Arqueologia e dos meios virtuais, não leva apenas ao desenvolvimento de novas metodologias, mas também discussões teóricas (ZUBROW, 2006). A chamada Arqueologia Digital, surge dessa necessidade de desenvolver e estreitar as relações entre uma disciplina essencialmente social com a exatidão da computação. Segundo Torres , este ramo da Arqueologia é conceituado como a capacidade de diferentes técnicas e/ou tecnologias de serem aplicadas às diversas etapas dos trabalhos arqueológicos.

A Arqueologia Digital tem um apelo técnico bem característico, por esta ser associada diretamente a tecnologia, equipamentos e ao maquinário, em geral, que auxiliam o arqueólogo no registro de campo. Segundo Hugget , a Arqueologia Digital tem se mostrado cada vez menos como apenas um apanhado de conhecimentos emprestados de outras ciências e mais como capaz de desenvolver novas técnicas e ferramentas que sirvam não só para uso na Arqueologia, mas como em outros campos científicos.

É fundamental, para que esse desenvolvimento ocorra, que haja discussões teóricas sobre o assunto, ou seja, a Arqueologia Digital precisa desenvolver sua base de ideias nos grandes eventos acadêmicos e até ser incluída nos currículos escolares em cursos de Arqueologia pelo mundo. Isso será importante, até para desmistificação de certos aspectos ligados a utilização de técnicas e tecnologias, as quais podem ser usadas de maneira errônea e causar uma sensação de um desenvolvimento científico baseado apenas na técnica e levar a danos irreparáveis ao patrimônio arqueológico .

Conhecer o equipamento com que se trabalha também é importante para evitar problemas, tanto os de ordem operacional, que podem resultar numa obtenção e gestão errada de informações, mas também os chamados “apertadores de botões”, ou seja, pessoas apenas com a capacidade de operar determinada ferramenta, mas que não sabem o que está por trás do processo em que estão realizando.

A compreensão destas questões fornece uma base para repensar a Arqueologia, como coloca Huggett (2017), quando define os chamados “artefatos cognitivos”, que podem ser desde uma colher de arqueólogo até uma estação total ou um computador, e tem a função de complementar a cognição arqueológica, possibilitando o pesquisador a enxergar elementos que seriam impossíveis apenas com o corpo humano.

Ao olharmos a Arqueologia Digital de um ponto de vista mais específico, podemos dividi-la em duas vertentes, uma a Arqueologia Virtual e a outra a Ciber-Arqueologia, sendo a primeira voltada para estudos que envolvem um sistema de dados estáticos e não modificáveis, já a segunda é justamente o oposto e trabalha com dados dinâmicos, capazes de gerar ambientes virtuais que simulam a realidade, em outras palavras podemos dizer que em um são feitas reconstruções estáticas do passado e na outra simulações, respectivamente .

Esta pesquisa, estará no campo da Arqueologia Virtual, já que trabalhará com sistemas de dados estáticos aplicados ao uso de um Sistema de Informação Geográfica (SIG). Segundo Conolly e Lake , o SIG é “uma plataforma para obter e integrar arquivos de dados espaciais, como mapas topográficos, localização e morfologia dos sítios, fotografia aérea (...) tudo isso em um meio analítico comum.” .

Os SIG lidam com dados estáticos, que podem ser divididos em dois grupos: os vetores, que são informações estruturadas em pontos, linhas e polígonos, obtidos mediante uma vetorização ou redesenho de um determinado arquivo; e os rásters formado por elementos compostos por *pixels*, pequenas matrizes de cores diferentes que dão forma a uma imagem .

2.1.1 Os Modelos Preditivos

O avanço tecnológico que fez dos SIG uma ferramenta essencial para análises espaciais e proporcionou a coleta de mais informações às pesquisas arqueológicas, contribuíram também para o desenvolvimento de metodologias analíticas com o uso de modelagens matemáticas e estatísticas.

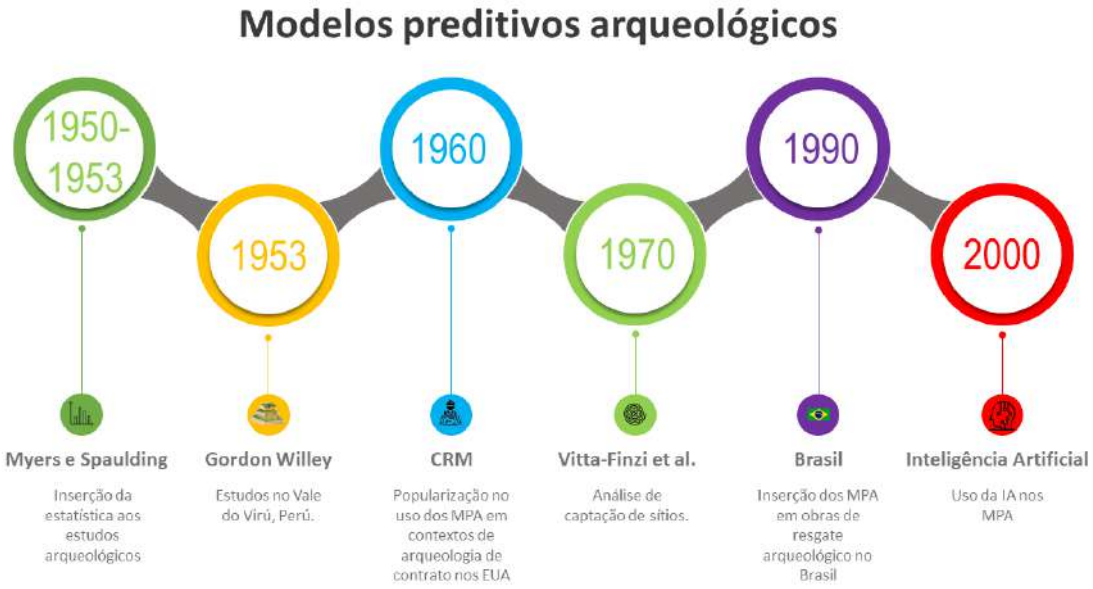
Os modelos preditivos são uma dessas metodologias que vem sendo incorporadas às pesquisas arqueológicas, sobretudo a partir de meados do século XX. A história do uso desses modelos, de um modo geral, está ligada ao início da era dos computadores, por volta da década de 40. À medida que os computadores iam evoluindo em suas capacidades de processamento, torna-se possível a aplicação de modelos computacionais mais complexos. Com isso, órgãos governamentais são os primeiros a se utilizar de análises de dados como forma de realizar previsões .

Como os computadores têm avançado constantemente desde o período acima citado, a discussão em torno de uma definição a respeito da modelagem preditiva está em constante desenvolvimento. Segundo Kunh e Johnson , o modelo preditivo é “o processo de desenvolvimento de uma ferramenta ou modelo matemático que gera uma predição acurada.”

2.1.2 Os Modelos Preditivos em Arqueologia (MPA): uma linha do tempo

O uso de modelos preditivos acompanha a Arqueologia desde antes das discussões trazidas em se estabelecer novas bases teóricas para a Arqueologia, com uma maior interdisciplinaridade. Sítios arqueológicos já eram buscados com base ao menos em aspectos lógicos, previamente conhecidos. Dito isso é importante relatar aqui o histórico do uso dessa metodologia, a qual segue a linha do tempo abaixo (**Figura 1**).

Figura 1: Breve linha do tempo dos estudos com MPA.



Fonte: Elaborado pelo próprio autor (2022)

O primeiro uso de um MPA remonta a década de 50, com os trabalhos do arqueólogo Gordon Willey , no Peru, com o intuito de compreender a distribuição espacial dos assentamentos pré-históricos no Vale do Virú. É apenas com Vita-Finzi e colaboradores , porém, que o uso dos MPA desenvolvem uma base metodológica mais específica. O desenvolvimento da noção de “análise de captação de sítios” pelos autores acima, buscava compreender a respeito da exploração de recursos por parte dos grupos humanos antigos, utilizando análises espaciais em grandes zonas que partiam de um centro específico.

Um pouco antes da publicação acerca da análise de captação de sítios, ainda na década de 60, os MPA começam a se popularizar, sobretudo no contexto da CRM (*Cultural Resource Management*), que compreende os trabalhos de Arqueologia de Contrato do cenário estadunidense. O crescimento acelerado do país da América do Norte, aliado a leis mais rígidas na defesa do patrimônio arqueológico, fizeram dos MPA um grande auxiliar .

Em linhas gerais, segundo Verhagen , a definição acerca dos MPA os descreve como meios de predizer onde estão materiais e sítios arqueológicos, baseados em amostras de dados já conhecidas ou a partir do entendimento de padrões de comportamento humano.

No Brasil, o cenário para a implementação de modelos preditivos, não é muito diferente do relatado para América do Norte, e começa a ser discutido justamente nesse período de grande aumento nas pesquisas para implantação de obras de grande impacto. Um artigo seminal para a discussão de tal tema foi escrito por Kipnis , onde ele versa sobre a importância dos modelos preditivos de modo que estes podem oferecer subsídios para a realização de planejamentos e nas questões referentes aos custo-eficiência em um trabalho de estudo de impacto ambiental no Brasil.

Aliado ao contexto infraestrutural, a Arqueologia também desenvolvia bases teóricas alinhadas a um pensamento hipotético-dedutivo, buscando trazer mais empirismo aos trabalhos, e em consonância a Teoria dos Sistemas. Essas características da Arqueologia Processual, tornam as pesquisas arqueológicas mais interdisciplinares e aspectos trazidos de ciências como a Geografia, a Matemática e a Estatística passam a ser mais comuns aos estudos .

Uma das figuras mais importantes nesse processo de fazer a Arqueologia dialogar com outras ciências foi David Clarke. Mesmo sendo contemporâneo aos estudos processualistas e se enquadrar como tal, o autor foi bastante crítico a forma como alguns pesquisadores relacionavam os aspectos arqueológicos e ambientais, por exemplo. Segundo o autor era preciso analisar o espaço arqueológico em diversos compartimentos e seguindo não só os aspectos ambientais, mas todo o contexto social no qual aquele material foi encontrado .

A Estatística também foi importante no processo de desenvolvimento dos MPA. Ela é introduzida já na década de 50 como os trabalhos de Myers e Spaulding e é cada vez

mais utilizada para analisar os dados de origem ambiental e social atrelado aos vestígios arqueológicos, que apareciam cada vez mais em maior número .

Segundo Verhagen e Whitley existem dois tipos de modelos. Um se baseia na dedução, ou seja, a partir de alguns pressupostos teóricos e/ou em informações pré-concebidas ou pré-estabelecidas, como por exemplo, a necessidade de se hidratar ou se alimentar de um ser humano, seria usada como base para buscar sítios arqueológicos próximos a áreas com fontes de água. Já o segundo, se baseia na indução, ou seja, em fatos já conhecidos e testados, guiados, muitas vezes, por implementações estatísticas. Um exemplo destas modelagens pode ser visto quando vai se buscar novos sítios como base na existência de outros já conhecidos.

Dentro dos modelos indutivos podemos destacar ainda dois tipos de composições diferentes. Quando pensamos em Arqueologia, estes podem ser formados apenas de uma única classe, a qual representaria os sítios arqueológicos, e a partir da localização espacial de cada um destes, é calculada a estatística das variáveis ambientais e/ou sociais associadas a estes, de modo a extrair padrões nos dados, os quais servirão para classificar os locais ainda não conhecidos.

Por outro lado, os modelos indutivos podem ainda ser formados por duas classes. Além dos elementos que se referem aos sítios arqueológicos, haveria também áreas de ausência de vestígio, o que fará o modelo trabalhar na dicotomia ausência/presença (VERHAGEN; WHITLEY, 2020). Esta segunda configuração, será usada neste trabalho, afinal trabalharemos com os dados dos sítios arqueológicos de grafismos rupestres catalogados durante o Projeto de caracterização e documentação dos sítios rupestres no semiárido de Pernambuco, que também gerou informações a respeito dos locais sem vestígios de representações parietais, decorrentes do monitoramento de um aparelho de navegação por satélite (GNSS) que marcou os locais prospectados pela equipe .

2.1.3 Alertas no uso da modelagem preditiva

Na definição de MPA aqui apresentada, Verhagen (2018) toca num ponto interessante, pois ele fala que deve se buscar “padrões do comportamento humano”, ou seja, estas técnicas devem se preocupar com questões de cunho social além da análise ambiental por

si só. Uma das grandes críticas na literatura atrelado ao uso dessas metodologias se deve ao fato de se usar muitos elementos do meio ambiente como variáveis e deixar de lado os aspectos mais humanísticos, configurando um determinismo ambiental (VERHAGEN; WHITLEY, 2011).

Como forma de resolver o problema do determinismo ambiental, Verhagen e Whitley (2011), propõem o chamado modelo preditivo cognitivo, o qual se baseia em conceitos da Arqueologia Cognitiva e na Arqueologia Comportamental , buscando colocar o Homem como um ser que tem suas escolhas e vontades, trazendo os elementos individuais do comportamento humano, aliado a uma percepção do espaço que busca observar o ambiente de uma forma ampla e o registro arqueológico como elemento que pode ser submetido a agentes sin e pós deposicionais.

Os processos cognitivos trazidos por Verhagen e Whitley se baseiam na questão de tomadas de decisão, que podem, por sua vez, ser divido em dois grupos de variáveis, as diretas, as quais podem ser percebidas com os sentidos humanos e as indiretas, representadas pelas informações adquiridas através da experiência, especulação. Alia-se a isso, a noção de percepção espacial a qual o contexto cognitivo permite, já que muitos dos "sítios" definidos pelos arqueólogos, podem ter sido selecionados de uma forma subconsciente e, portanto, é necessário que a observação do sítio vá além da presença de materialidade, já que a paisagem e as unidades territoriais em que este(s) grupo(s) viviam é muito ampla.

Aqui usaremos a ideia do modelo preditivo cognitivo para inserir determinadas variáveis com foco socio-ambiental, tais como a captação de recursos, a proximidade para cursos d'água ou litologias mais comuns de serem feitos registros rupestres, ou ainda litologias relacionadas a fontes de ocre, por exemplo.

O outro problema associado ao uso de MPA, sobretudo quando relacionados a um contexto de salvamento, é o fato de se privilegiar áreas em detrimento de outras. Isso se deve primeiramente ao fato de os resultados de um modelo serem expressos por meio de uma probabilidade, ou seja, haverá locais com diferentes níveis de "chance" de se encontrar um vestígio arqueológico.

O segundo ponto está no fato de os modelos preditivos necessitarem de etapas de avaliação, sejam elas por meio da estatística, a qual medirá o quão bom seu modelo é face aos

dados fornecidos, mas também de uma avaliação observacional, em que a ida a campo e o teste *in loco* da performance do produto gerado, avalia não só os locais de alta probabilidade, mas também os que tem um nível médio ou baixo. Só com este tipo de estratégia é que se constrói modelagens mais coerentes, afinal de contas estes são apenas frutos de cálculos estatísticos que jamais podem ser levados como uma única solução de um determinado problema .

2.1.4 A gestão dos dados digitais

As críticas acerca dos modelos preditivos podem ser ampliadas ao olharmos para o trato com os dados digitais que compõe este ou qualquer outro tipo de modelagem matemática. Os avanços tecnológicos trouxeram uma gama de possibilidades na obtenção dos dados de uma pesquisa arqueológica. Essa revolução vem levantando debates acerca da chamada "era da Big Data" na Arqueologia. Em face disso, Huggett fala a respeito dos novos paradigmas que os estudos arqueológicos vem passando devido a esse crescente ganho de informações. Para o autor, a virtualização dos dados traz ganhos substanciais às pesquisas arqueológicas, pois torna as informações mais flexíveis, pesquisáveis, compartilháveis, calculáveis, tudo de uma forma mais rápida.

Por outro lado, há alguns cuidados que precisam ser levados em conta quando se lida com uma grande quantidade de dados digitais. A compreensão de que os dados são simplesmente meios para o fim, é um deles; Além deste, as reduções e compressões de arquivos, podem gerar perdas de informações, portanto saber lidar com isto é fundamental; alguns conhecimentos extras de computação podem ser exigidos também; e por fim, o uso deliberado das técnicas e tecnologias, pode ser um problema operacional grave, pois é preciso haver um propósito claro para a utilização de alguns equipamentos e metodologias nos trabalhos arqueológicos, caso contrário configurar-se-ia como uso pelo uso (HUGGETT, 2020).

Um outro grande problema ocasionado pelos dados é a sua manipulação deliberada com a intenção de mascarar uma realidade, ou seja, enviesá-los ao bel prazer. No clássico livro escrito por Darrel Huff, "Como mentir com estatística" , são levantados diversos exemplos de como é possível utilizar informações de forma tendenciosa, seja na Política, na Medicina e até nas Ciências Humanas, como forma de ludibriar a sociedade ou ter acesso a determinados privilégios.

Por mais que os exemplos do livro citado não sejam diretamente ligados à Arqueologia, eles servem como forma de alerta para mostrar os riscos que se há ao trabalhar com a estatística. Além disso, os modelos estatísticos, em qualquer área, são manipulados por um ser humano, que tem ideias próprias, que muitas vezes podem passar tanto para a composição do modelo, como para suas análises. Cabe, portanto, a cada pesquisador, perceber quanta interferência "externa" está sendo incluída em suas interpretações .

A manipulação errônea dos dados também pode causar danos sociais irreparáveis. As grandes empresas de tecnologias têm usado dos nossos dados, para comercializá-los, de modo que propagandas nos atinjam com o intuito de consumirmos determinado produto. Essa lógica um tanto quanto opressora, tem feito as pessoas se tornarem reféns da tecnologia, já que estas práticas vem moldando até nosso modo de vida .

A sociedade capitalista em que boa parte da população global está inserida, faz do uso de modelos uma forma de avaliar a sociedade de diversas maneiras. O uso de computadores tem criado "leis universais" que moldam os comportamentos humanos em detrimento do lucro das grandes empresas de tecnologias que desenvolvem estes sistemas .

Segundo O'Neil existem 3 elementos que envolvem os chamados Algoritmos de destruição em massa (ADM), que é esse tipo de modelo predatório relatado. Para a autora a opacidade, ou seja, a falta de transparência na divulgação de como cada tipo de avaliação é feita é um ponto; a escala é um segundo ponto trazido, no qual colocam esse tipo de prática a um dimensionamento tão amplo numa sociedade, que acabam criando "normas irrefutáveis"; e por fim ela causa danos, os quais podem ser irreparáveis para um indivíduo ou grupo, em detrimento de uma bonificação para uma seleta parte da sociedade.

Trazendo esse debate para Arqueologia, isso é tão perigoso quanto, pois o Patrimônio Cultural está sob ameaça, caso algumas técnicas, tais como as modelagens preditiva em Arqueologia sejam usadas de maneira incoerente, como na questão da avaliação observacional, a qual pode focar apenas em locais de alta probabilidade. Esta estratégia pode levar a danos irreparáveis ao patrimônio arqueológico.

Um outro fator que precisa ser discutido com relação a todo esse crescente acúmulo de informações no cenário dos dados digitais e no aprimoramento de técnicas e metodologias

nas ciências de um modo geral, é a “democratização dos dados”, ou seja, é o ato de tornar as bases de dados acessíveis a todos, de maneira gratuita. A falta de acesso às informações, um dos grandes problemas dentro da Arqueologia, e isto dificulta, inclusive, a troca de ideias entre os pesquisadores. Portanto, para que haja um desenvolvimento maior nas pesquisas arqueológicas, se faz necessário que os dados das pesquisas se tornem abertos. .

A abertura dos dados porém levanta um debate ético, já que em um editorial recente publicado pela revista *Nature* , ficou comprovado que os autores que não compartilhavam seus dados de forma integral, tendiam a ser inseridos em outras publicações relacionadas, já àqueles que abriam toda a sua base de dados, não ganhavam os mesmo méritos, sendo apenas citados em anotas de rodapé ou agradecimentos.

A divulgação desse tipo de informação, não passa apenas pelos dados brutos, os quais servem para construir a base estatística ou uma análise espacial dos trabalhos, mas também os próprios artigos, que ainda hoje são ligados a revistas que tem uma plataforma paga, as quais exigem do autor uma taxa para tornar uma publicação aberta ou o leitor precisa desembolsar uma quantia para poder ler tal obra . Para isso, plataformas como o *Research Gate* e o *Academia.edu*, tem se tornado grandes aliados na divulgação de trabalhos científicos de forma gratuita, no caso do primeiro ou parcialmente gratuito, no caso do segundo (TORRES, 2017).

O acesso aberto aos dados é de fundamental importância quando se fala em modelagens preditivas aplicadas à Arqueologia, já que estas trabalham com uma grande quantidade de informações que compõe as variáveis a serem usados nos cálculos estatísticos para obtenção dos produtos. Para a construção dos MPA, por exemplo, a presença de plataformas SIG abertas como o *QGIS* , integradas a gestores de bancos de dados como o *PostgreSQL* , facilitam a aglutinação dos dados espaciais com dados estatísticos.

A popularização de linguagens de programação de código aberto, tais como o *Python* e *R*, tem proporcionado uma evolução nas análises espaciais e estatísticas em Arqueologia também, e representam mais de 30% no mercado global da tecnologia da informação, segundo o *Popularity of Programming Language Index* (PYPL, 2021), que é medido através das buscas pelos nomes das linguagens na plataforma Google. Boa parte da aplicação destas linguagens está atrelado à Ciência dos Dados, que como nome já sugere, se

preocupa com análises e o entendimento de bases de dados, relacionadas a fenômenos humanos, sociais e naturais .

Estas duas linguagens de programação supracitadas têm bibliotecas especializadas para auxiliar os utilizadores a criarem as estatísticas de seus modelos. O *Python*, por exemplo, tem inúmeras bibliotecas que podem ser usadas para a análise de dados e/ou desenvolvimento de modelos de aprendizagem de máquina, tais como o *Scikitlearn, Tensorflow, Keras e PyTorch*, além destas, há outras auxiliares, como o *Pandas, Matplotlib e o Numpy* (GÉRON, 2019).

Dentro da linguagem *R*, que é um projeto especializado para o desenvolvimento de modelos estatísticos e na análise dos dados, as bibliotecas acabam sendo muito diversas, pois existem, muitas vezes pacotes bem específicos, para cada determinado tipo de algoritmo, porém um dos mais usados são o *CARET, alrr3, lmtest, arules*. Além destes, um grande suporte, a qualquer análise de dados, é a geração de gráficos, que no *R*, tem na biblioteca *ggplot2*, seu grande trunfo .

Entre essas duas linguagens trazidas, o *R* ainda tem bastante destaque dentro das Ciências Humanas e consequentemente, na Arqueologia. Segundo uma pesquisa feita por Schmidt e Marwick (2020), existem diversas revistas de Arqueologia onde é possível ver a citação a tal linguagem, e ao longo dos anos, a inserção deste assunto vem sendo ampliada. Um outro ponto observado pelos autores é que boa parte dos artigos que usam do *R* estão ligados a estudos que envolvem uma análise ambiental, o que faz sentido, justamente pelo fato desta possuir uma integração com o QGIS, que é uma plataforma gratuita e oferece as mesmas possibilidades dos exemplares pagos.

Ainda no âmbito da discussão a respeito da abertura de dados, as linguagens de programação e plataformas como o *Github*, favorecem a divulgação e troca de informações através de códigos e/ou modelos aplicados a Arqueologia , mas estes ainda precisam ser mais utilizados, pois ainda hoje, a quantidade de pesquisadores de Arqueologia que usam este ambiente digital, é muito pequena.

Compreender os riscos atrelados aos dados digitais e aos próprios MPA, aliados a um acesso aberto e compartilhado de dados das diversas pesquisas realizadas na Arqueologia,

dão subsídios ao desenvolvimento de grandes redes de informação, que poderão ser usadas no desenvolvimento de metodologias para melhor compreender a relação entre os grupos humanos pretéritos e o meio ambiente. A . Além disso, a aglutinação e boa gestão dos dados digitais permite que a Arqueologia se conecte a uma área da computação que demanda grandes bancos de informações digitais, que é a Inteligência Artificial (IA).

2.2 A Inteligência Artificial

A Inteligência Artificial (IA) e mais precisamente a Aprendizagem de Máquina (AM), aos poucos vem ganhando espaço nos estudos arqueológicos. Neste trabalho em específico serão usados alguns algoritmos computacionais em conjunto com a aplicação dos MPA.

O termo algoritmo irá aparecer ao longo deste trabalho diversas vezes, e para tanto precisa ser conceituado, apesar do seu uso ser bastante comum ultimamente. Segundo Cormen e colaboradores , estes são uma sequência finita de etapas computacionais que transformam os dados de entrada em dados de saída ou ainda uma ferramenta para resolver problemas computacionais específicos.

Além do algoritmo, é importante aclarar alguns aspectos a respeito da IA também, afinal esta não é uma técnica que irá criar produtos inteligentes do nada, como se vê em filmes, e para que a mesma funcione é preciso que um operador forneça dados e cálculos a uma máquina. Segundo Russel e Norvig , podemos defini-la como uma técnica que pode fazer uma máquina para agir ou pensar como seres humanos

Para que haja essa reprodução, porém, é necessário a construção de uma rede de informações que será treinada, para que o computador aprenda com esses dados e a posteriori execute o aprendizado. Este conceito advém de estudos muito antigos, desde a Grécia Antiga, quando os filósofos, perceberam que o conhecimento humano está atrelado a códigos de linguagens que definem quais ações um indivíduo toma .

Claramente a IA não se baseia apenas nos pensamentos da Filosofia, mas também em outras áreas, como a matemática, a qual fornece a base lógica para compreensão dos cálculos referentes aos algoritmos; a economia, no sentido de buscar alternativas mais rentáveis na solução dos problemas; a Neurociência, no sentido de compreender como se dão as relações

neuronais nos cérebros humanos; a Psicologia, para compreender como humanos e animais pensam e agem; e a engenharia computacional, no sentido de desenvolver um maquinário eficiente para desenvolvimento de aplicações em IA .

A estatística, que também é essencial a IA, sobretudo no que diz respeito ao desenvolvimento das aplicações matemáticas, tais como a regressão linear, que foi pautada ainda no século 19, em artigos escritos por Legendre e Gauss, serviu de base primeiramente a estudos na Astronomia. Nos anos de 1940, vários autores trabalharam sobre os modelos lineares e desenvolveram a chamada regressão logística. Todas essas implementações, estavam, porém, no campo dos modelos lineares, é só a partir da década de 80, com o avanço computacional, que os chamados modelos não lineares são postos em prática .

O desenvolvimento da estatística proporcionou a base para a criação dos modelos de aprendizagem, e com isso, a primeira modelagem a se utilizar de uma, ainda incipiente, técnica da AM, foi produzido nos anos 40, por McCulloch e Pitts, um neurofisiologista e um matemático, respectivamente, se aproveitando de um momento em que as pesquisas em neurociências e na implementação de técnicas estatísticas mais robustas se aprimoravam, produziram um modelo que usava um circuito elétrico para mostrar como se dava o comportamento dos neurônios. Esse fato é considerado o início das redes neuronais artificiais, uma das técnicas do AM .

A aplicação em AM mais conhecida, no entanto, só é reproduzida em 1950, quando Alan Turing, desenvolve o *Turing Test*, que colocava homem e máquina frente a frente. Uma pessoa escrevia uma pergunta a mão para um computador e após isso recebia respostas, o computador por sua vez, passaria no teste a partir do momento em que não fosse possível distinguir quando a resposta enviada de volta tivesse sido produzida pela máquina ou por uma outra pessoa .

Em sequência ao estudo de Turing, novas pesquisas começam a surgir, dentre elas o desenvolvimento da primeira máquina a rodar um sistema de redes neuronais artificiais, o chamado *Perceptron*, desenvolvido por Frank Rosenblatt (1952), da Cornell Aeronautical Laboratory .

Com esse histórico anteriormente apresentado, fica evidente que a evolução da IA ou mesmo do AM, está em alinhamento com o próprio entendimento de como o cérebro humano funciona e como o comportamento humano influência nas ações que são tomadas durante o cotidiano. Isso mostra o quanto as tecnologias computacionais, são mais próximas de ciências humanas do que se pode imaginar. As máquinas apenas buscam, de alguma forma, representar e achar padrões de comportamento humano.

Segundo Burkov (2019), "as máquinas não aprendem", mas simplesmente nos dão respostas baseados em dados estatísticos que serão testados de acordo com um método adequado para tal situação. Ainda de acordo com o mesmo autor, o Termo "Aprendizagem de Máquina" só recebeu esse nome no fim da década de 50, como uma jogada de marketing da IBM[1].

No sentido geral, podemos conceituar o AM como a ciência capaz de programar computadores a aprender a partir de um conjunto de dados . De uma forma mais específica, ela é tida como um subcampo da Ciência da Computação que lida com a construção de algoritmos, baseados em uma coleção de dados provenientes da natureza, do homem ou de outros algoritmos . Em complemento a esta definição, Kelleher, Mac Namee e D'Arcy , o conceituam como um processo automatizado em que se extrai padrões através dos dados.

Atualmente o AM tem sido usado em diversas atividades cotidianas, que muitas vezes nem nos damos conta. As câmeras dos celulares, por exemplo, contam com sistemas inteligentes, para melhorar a qualidade das fotos tiradas. Na segurança pública, as câmeras podem ser usadas na identificação facial de criminosos; no mercado financeiro, o AM pode traçar cenários econômicos e até auxiliar o investidor a realizar suas aplicações. Os veículos autônomos também são outra realidade que essa tecnologia proporciona .

As possibilidades do AM são inúmeras, e dentro das ciências sociais e da terra, elas já começam a ser aplicadas. Um exemplo de aplicação nas ciências da terra é o trabalho de Rahmati et al , que faz uma avaliação de diferentes modelos de AM, para identificar locais suscetíveis a deslizamento de encostas. No campo das ciências humanas, até mesmo a Filosofia,

[1] International Business Machines Corporation – Empresa americana do ramo tecnológico

se apropria da tecnologia, podendo utilizar-se dos modelos computacionais para testes sobre a linguagem, pensamentos e metodologias .

2.1.1 A Arqueologia e a IA

Dentro do contexto da modelagem preditiva de um modo geral, a IA surge como uma forma de diminuir o enviesamento atrelado aos MPA. A partir do fim dos anos 1960, com a Nova Arqueologia, é que a estatística e a computação começam a ser vistas como formas de auxiliar o arqueólogo nos processos de experimentação, simulação e análises gráficas, o que contribui para a diminuição da subjetividade nos estudos analíticos em Arqueologia .

Os “Novos Arqueólogos” se utilizaram da Teoria dos Sistemas, e observavam a sociedade humana como uma máquina, estas conclusões foram alvos de grandes críticas, pois tentavam limitar o comportamento humano a teorias universais simples. A partir dos anos 80 porém, essa visão é mudada com os novos paradigmas da IA, o conhecimento passa a ser visto como construído através da percepção de ações que limitam ou guiam o comportamento humano .

Com isso em mente, a AM aparece na Arqueologia com a capacidade de fazer as máquinas aprenderem com as inferências e erros que são postos a ela. Esse acúmulo de informação e o aprendizado partir disto, faz com que os modelos de aprendizagem de máquinas deem respostas mais graduadas a Arqueologia, permitindo que o pesquisador compreenda melhor as limitações e ganhos do comportamento humano .

2.1.2 Aplicações na Arqueologia

Os algoritmos de AM têm aplicações diversas dentro da Arqueologia, que podem perpassar por todos os tipos de aprendizagem citados. A classificação tipológica de vestígios arqueológicos é o ramo mais explorado, no quesito de uso automatizado de processos nas Ciências Arqueológicas . Um grande exemplo de aplicações nesse ramo classificatório-tipológico é o projeto ArchAIDE, que visa identificar fragmentos cerâmicos a partir de uma

foto . Flores et al , também desenvolveram um sistema de classificação, porém para pontas de projéteis.

Dentro do estudo de práticas para o achado de novos sítios arqueológicos e na aplicação em MPA, podemos destacar alguns trabalhos que se utilizam do AM, como é o caso de Caspari e Crespo (2019), Kramer (2020), Cowley et al. que buscaram identificar, de maneira automatizada, em imagens de satélites, marcas nos solos que indiquem a presença de sítios arqueológicos. Há também pesquisas no sentido de criar modelagens que extraem de imagens de satélites os aspectos ambientais em forma de valores estatísticos para o encontrar os sítios arqueológicos .

No que diz respeito a utilização da AM e da própria estatística em sítios com grafismos rupestres podemos destacar uma série de trabalhos focados não só no uso de MPA , mas também em análises micro espaciais, utilizando cálculos estatísticos para agrupar elementos pictóricos de determinados sítios (MATOS, 2019).

2.3 O Sítio Arqueológico

De nada adianta o uso de todas estas técnicas mencionadas, se não temos um arcabouço conceitual. Esta tese antes de tudo trata a respeito das análises espaciais em Arqueologia. A busca para compreender as relações entre os grupos humanos do passado e o meio ambiente em que estes estavam inseridos é um debate que há muito tempo já ocorre. Desde o período dos colecionistas, no século XVI, a noção de escala e análises espaciais em níveis locais foram realizadas em atividades de campo. A cartografia também foi inserida às primeiras pesquisas arqueológicas aplicadas em trabalhos na Grã Bretanha, por exemplo, como auxílio na localização de sítios arqueológicos .

É a partir do século XX, porém, que a discussão em torno do estudo do espaço ganha mais destaque. Os trabalhos de Grahame Clark e Julian Steward , por mais que tivessem um enfoque ainda muito descritivo, são um marco significativo para os estudos espaciais na Arqueologia, principalmente pela discussão trazida. Clark atrelava a mudança cultural a questões de instabilidade temporal, devido a modificações ambientais e nas taxas populacionais, inovações e no contato. Já Steward, que trabalhava na perspectiva de uma

“evolução cultural”, observou que as variações culturais só seriam perceptíveis através de uma análise a respeito da adaptação do ser humano ao meio.

Grande parte das discussões iniciadas no século XX acerca dos estudos espaciais carregavam uma base da Geografia. Para tanto, alguns conceitos bases advindos da Ciência da Terra precisam ser aqui elencados. Para exemplificar, partiremos do que a Sociedade Real de Geografia chama de “conceitos de alto nível”, dos quais este trabalho se voltará a dois destes: o espaço e a escala.

Partindo primeiramente de uma visão geral acerca do conceito de espaço na Geografia Humana, Santos o coloca como um local onde homem e meio se relacionam de forma dinâmica, ou ainda, um sistema de valores, o qual está em constante transformação no que diz respeito a função e significado dos seus objetos.

Em complemento a definição acima, é possível afirmar que o espaço é antes de tudo uma estrutura social que guarda aspectos intrínsecos da sociedade, como as relações de poder, os sistemas políticos e todos os tipos de inter-relações entre o homem e o meio .

O segundo conceito trazido como de “alto nível” na Geografia, é o de escala. Este pode ser definido como a representação de uma área, produzida e definida por processos sociais, desde uma pequena unidade, o corpo, até a maior delas, o universo. (CAMPBELL, 2016).

Na Arqueologia, os trabalhos em análise espacial e os conceitos de espaço, têm como base fundamental os estudos de geoantropólogos, como Ratzel, Frobenius e Gradmann, (ROOD, 1982). A utilização do pensamento destes pesquisadores nos estudos arqueológicos na Europa e EUA se deu de formas diferentes. Enquanto no continente europeu o espaço foi utilizado com uma dimensão geográfica, na América do Norte, a busca por padrões de assentamento e organização social, fez com que os estudos espaciais tivessem um viés mais antropológico.

A escala nos estudos espaciais e mais especificamente nos trabalhos preditivos é de extrema importância. É sabido, porém, que haverá limitações na extração de alguns padrões ambientais associados, pois há elementos que demandam uma resolução espacial maior ou menor para serem percebidos, afinal é impossível se ver todos os elementos que compõe a paisagem a partir de um único nível de visão. Além disso a escala temporal é outro fator

importante, que aqui terá que ser preterido, afinal o objeto de estudo desta pesquisa é um vestígio impossível de ser datado diretamente e com poucas ou nenhuma cronologia relativa associada e aqui será considerado como um elemento neutro .

O uso das escalas em Arqueologia passa por uma série de classificações de acordo com os níveis de observações diferentes. Clarke , por exemplo, definiu a chamada Arqueologia Espacial como “o estudo do fluxo e a integração das atividades humanas, dentro e entre estruturas e recursos espaciais, em diferentes níveis de percepção”.

Esses diferentes níveis de percepção postos na definição acima, caracterizam as diversas escalas de análise definidas por Clarke (1977). Nesta pesquisa, o enfoque visa observar o ambiente arqueológico de uma perspectiva mais ampla, a nível regional e visualizando as relações extra-sítio ocorridas na bacia do Pajeú, sendo denominado de “escala macro”. Com a mesma nomenclatura e em complemento ao que prega a Arqueologia Espacial, a visão da Arqueologia Ambiental, representada aqui por Dincauze , coloca ainda que esta abrangência de escala busca compreender as relações de comunicação e exploração de recursos pelos grupos humanos.

Um outro detalhe referente às escalas está na representação destas nos meios digitais, tais como o próprio SIG, que aqui será uma ferramenta fundamental. Segundo Lock e Pouncett , o SIG tem por condicionante a utilização de uma “escala analítica”, que vai buscar, medir e analisar padrões no espaço arqueológico, por outro lado, a escala fenomenológica, que visa buscar as interações e experiencias cotidianas no mundo, precisa também ser integrada às pesquisas espaciais em ambientes digitais, para que se tenham análises mais completas. Essa discussão vai ao encontro das críticas apontadas ao MPA, no sentido de haver um determinismo ambiental atrelado aos estudos deste tipo de metodologia.

Dentro desta pesquisa, se buscará trazer os elementos socio-ambientais, os quais podem abarcar a escala fenomenológica, mas que não será um objetivo direto deste trabalho. O ambiente, porém, é essencial aqui, no sentido metodológico, até porque o início da construção de um MPA, passa pela delimitação de uma área que será conceituada como um sítio arqueológico, o qual é o alvo deste trabalho.

O sítio arqueológico, porém, está diretamente associado ao chamado espaço arqueológico. Este pode ser composto de 5 características, dentre as quais: é ligado ao tempo; está relacionado com as práticas diárias de um grupo; é difícil de ser representado; trata muito mais das ausências do que das presenças; e é atrelado a mobilidade (GILLINGS; HACIGUZELLER; LOCK, 2019).

Destacando sobretudo as duas últimas características acima, podemos dizer que as ausências, partem do ponto de que os vestígios arqueológicos e o espaço onde estes estão inseridos, não se limitam apenas a área definida convencionalmente como "sítio arqueológico", pois, com toda certeza, locais em que não há vestígios, mas ficam em áreas periféricas próximas de concentrações, certamente tratam-se de ambientes em que os grupos humanos do passado estiveram, afinal o espaço é antes de tudo atrelado a mobilidade, e os grupos humanos precisavam realizar suas atividades de subsistência e para tal se deslocavam .

Desde a década de 70, Plog, Plog e Wait já enxergavam essa noção não limitante para os sítios arqueológicos. Eles denominaram de "não-sítio" os locais onde há presença de atividade humana, porém não é possível delimitá-los espacialmente. Esta visão foi corroborada pelos estudos de Schiffer a respeito da Formação do registro arqueológico, quando é lançado luz aos artefatos descartados e ao reuso, indicando essa dinamicidade nos sítios arqueológicos, e se opondo a visão materialista e estática de figuras como Willey e Phillip e Binford , que viam a necessidade de haver uma concentração vestigial para denominar o espaço de sítio arqueológico.

Todo esse debate em torno do conceito de sítio arqueológico será muito importante para a delimitação de cada um dos sítios de grafismos rupestres da Bacia do Pajeú. Os aportes metodológicos desta pesquisa, fazem com que o conceito de sítio arqueológico aqui utilizado tenha uma noção materialista, já que é necessário fornecer informações mensuráveis às ferramentas.

Como aqui teremos além da classe que representa os sítios, uma outra categoria que representam as ausências, estas serão aqui entendidas como locais que foram visitados durante uma prospecção, mas que não se achou vestígio rupestre algum . Sabe-se, porém, de acordo com a própria noção do sítio arqueológico como um elemento mais amplo no espaço, que

eventos pós-deposicionais, ou o simples fato de um grupo passar pelo local e decidir não deixar nenhum rastro, podem provocar ausências vestigiais .

Apesar da limitação imposta pela metodologia, que exige uma delimitação espacial exata, isso não irá impedir que nas considerações e interpretações realizadas ao longo do trabalho, a noção mais ampla de sítio, que insere os grupos humanos dentro de um ambiente mais vasto de possibilidades, seja levada em consideração. Este tipo de análise, pode levar a outras interpretações que não seriam possíveis quando se pensa em um espaço arqueológico estático.

2.4 Grafismos Rupestres

Os registros rupestres podem ser classificados como um conjunto pictórico confeccionados sob uma camada rochosa fixa em um determinado local. Pode ainda ser considerado, dentro da Arqueologia e Antropologia, como um marco testemunho da ocupação humana .

Um dos principais elementos estudados nos últimos anos relativos aos grafismos rupestres são a localização e distribuição destes vestígios, desde uma escala micro , onde se levam em conta aspectos específicos dos registros de um determinado sítio arqueológico, ou ainda em uma escala macro , observando os elementos ambientais que possam ter sido levado em conta para que os grupos pré-históricos escolhessem determinado local para fazer uma representação parietal.

As escolhas dos locais onde seriam feitos os desenhos passavam não só por uma perspectiva ambiental, mas estavam principalmente associados a questões simbólicas, impossíveis de serem desassociadas do comportamento humano . Isto faz com que a localização de um sítio, seja ele de pintura ou de gravura , passe a ser um elemento ativo na relação com a simbologia .

Os padrões da distribuição espacial de sítios arqueológicos com grafismos rupestres podem estar relacionados a identidades étnicas de acordo com Pillay, O’regan e Emmitt . A maneira como os grupos pré-históricos e mesmo as sociedades atuais interagem com o meio em que habitam é guiada pelos aspectos culturais.

Os grafismos rupestres são o vestígio mais estudado na Arqueologia do Nordeste, afinal eles estão distribuídos por toda a Região. A dinâmica por trás destes registros parietais vem sendo estudada desde a década de 70. Muitas das teorias ainda estão sendo discutidas nos dias atuais. Martin e Guidon , por exemplo, já falavam da necessidade de buscar novos olhares para entender a dispersão deste tipo de vestígio humano.

As diversas tradições rupestres existentes, tem cada uma, características específicas quanto a técnica e estilo. Além disso, acredita-se que cada uma delas teve um epicentro de surgimento. A Tradição Nordeste, por exemplo, teve sua possível origem dos grupos que vieram para o sudeste piauiense, mais especificamente para o PARNA Serra da Capivara, margeando os contornos do Rio Parnaíba. Essas conjecturas referentes a distribuição espacial são limitadas, até por que não há cronologias suficientes, para corroborar este tipo de hipótese .

O que se observa, porém, com as análises mais específicas destes grafismos rupestres da região da Serra da Capivara, é que há semelhanças com outros desenhos encontrados em três áreas de difusão distintas: o Seridó (RN/PB), o PARNA do Catimbau (PE), e o Vale do São Francisco desde Sergipe até o Norte de Minas Gerais, passando pela área de Central, na Bahia. Os chamados grafismos emblemáticos, como as figuras antropomórficas costa contra costa ou ainda os cabeças de caju, são os melhores exemplos dessas recorrências. .

Além dos grafismos da Tradição Nordeste, há outros grupos bem característicos na Região Nordeste do país, dentre os quais a Tradição Agreste, que também está muito difundida, sobretudo na Região do Pajeú. Esta pode ser caracterizada por seus grafismos puros, desenhos isolados e marcas de mão. Essa última característica inclusive, é um elemento que pode ser encontrado em várias regiões do Nordeste, exceto na Serra da Capivara, o que indica que esta técnica deve ter se originado na área central de Pernambuco e se expandido para outros locais, usando sobretudo o vale do São Francisco como uma rota de difusão .

Além de pinturas, o Pajeú também guarda muitos sítios de gravura. Não há ainda uma unanimidade na academia que separem as diferentes gravuras em grupos, sendo a Tradição Itacoatiara, a única tentativa de agrupamento para este tipo de técnica. Esta, porém, foi delimitada com base em sítios como a Pedra do Ingá, que tem representações bem características que não são recorrentes em todos outros sítios gravados .

Os estudos referentes às gravuras rupestres tem indicado, porém, que essas representações tem uma forte compatibilidade com cursos d'água , sugerindo uma relação simbólica ou ainda uma necessidade dos grupos pré-históricos em utilizar o recurso hídrico para aplicação da técnica .

No que diz respeito ao uso de implementações estatísticas no ramo dos grafismos rupestres da região Nordeste, podemos destacar dois trabalhos distintos, cada qual usando metodologias únicas. Van Havre , que se utiliza da análise das redes e de correspondências múltiplas para se chegar a marcadores de identidade ou padrões gráficos dos grafismos rupestres.

O outro trabalho, conduzido por Matos (2019), em que também se observam aspectos particulares dos grafismos, e se utilizam também de técnicas estatísticas, como a análise de componente principal (ACP), de modo a criar grupos distintos de desenhos, a partir de variáveis morfológicas. Foi usado também a análise cladística, que é um método praticado na Biologia para reconstruir processos evolutivos e na análise de grafismos rupestres tem a função de observar relações entre os padrões gráficos das representações. Além disso, o autor usa da análise locacional, que tem o intuito de relacionar as variáveis ambientais com os locais em que os sítios estão inseridos.

Esta utilização de recursos estatísticos na Arqueologia e em especial nos estudos do registro rupestre, mostram o potencial deste tipo de recurso. O estudo que será aplicado ao Pajeú, terá a intenção de ampliar este leque de possibilidades já existentes e ao mesmo tempo procurar extrair ainda mais informações a respeito dos grafismos rupestres do Nordeste e mais ainda sobre os grupos humanos que por ali passaram.

Padrões geoambientais

O principal interesse deste trabalho é entender a respeito dos padrões ambientais associados aos sítios arqueológicos. Aqui teremos como referência os locais com registros rupestres, afinal são os exemplares achados em maior número dentro dos limites do Pajeú. Por nos atermos a uma escala de abrangência macro, será, à primeira vista, impossível trabalhar com aspectos específicos a respeito das manchas gráficas em si, isto, porém, não quer dizer que não será necessário deixar claro algumas questões a respeito deste tipo de vestígio arqueológico.

2.4.1 O caráter comunicativo dos grafismos rupestres

A relação ambiental do registro rupestre, passa também por um aspecto de visibilidade, que por sua vez coloca esse tipo de representação dentro de uma perspectiva comunicativa, afinal se boa parte dos grafismos eram feitos para serem vistos, alguma mensagem podia ser compreendida a partir deles . O próprio Pajeú tem exemplares de sítios que ficam em locais visíveis a quilômetros de distância.

Sanchidrián , aponta que a comunicação verbal foi a base para a evolução da nossa espécie, hoje sabe-se, que inclusive o *Homo erectus* se utilizava da voz para se relacionar, mas é com os Homo *sapiens* que a verbalização se torna muito mais comum. Diante disso, é sabido que grupos diferentes também poderiam ter línguas diferentes, e a comunicação pela voz seria prejudicada, para tanto, os grafismos poderiam aparecer como uma forma de entendimento entre as partes.

A comunicação pode ser vista dentro do ambiente arqueológico na diversidade dos registros gráficos, na frequência com que estes aparecem no espaço, e na diversidade de cores e técnicas empregadas para se fazer um grafismo (TRONCOSO, MOYA E BASILE, 2016; MATOS, 2019). Em uma análise de grafismos rupestres em duas regiões diferentes no Chile, Troncos, Moya e Basile (2016), observaram que havia diferenças claras nas representações entre às áreas de estudo. Cada área apresentou um “horizonte compartilhado” para as comunidades que ali viviam, os integrando as mesmas práticas e a processos similares de significação do espaço.

Numa escala mais regional, nesta mesma pesquisa, Troncoso Moya e Basile , identificaram ser possível ver algumas recorrências entre duas diferentes áreas do espaço, porém, quando a escala aumenta a sua resolução, focando em um território apenas, observa-se que em um determinado local da área de estudo, se tem uma produção e um fluxo muito maior de grafismos, e consequentemente de informações, já que esses locais apresentavam uma densidade demográfica, uma disputa por territórios e busca por alianças maior do que outras regiões.

Para se entender, porém, os elementos comunicativos de um registro rupestre, alguns procedimentos podem ser adotados e dentre eles podemos destacar a busca pela identidade cultural. As representações pré-históricas tem particularidades que podem ser reduzidas à identidades, que na verdade são elementos característicos usados para diferenciar ou reconhecer um grupo de representações, que se configuram como uma identidade cultural .

O processo de evolução das identidades gráficas é diretamente dependente de aspectos espaciais e temporais e perpassam por análises visuais que os definem como de natureza material e que portanto tem uma estrutura física, seja ele uma gravura ou uma pintura, mas também há elementos não-materiais que precisam ser levados em conta, tais como às práticas sociais e o proceder técnico de manufatura dos grafismos .

As identidades observadas no registro, não se resumem às práticas de pintura e gravura executada pelos povos pré-históricos e segundo Van Havre estes já faziam parte das interações entre os indivíduos, que podiam ter sido expressas em diversos tipos de vestígios arqueológicos, inclusive na pintura.

Em resumo, é possível se identificar 3 tipos de escalas dos registros rupestres. Um tem a escala regional, que guiará boa parte deste estudo, preocupada com a relação de uma série de sítios e os aspectos ambientais e sociais associados a estes. A segunda escala é a escala de sítio, a qual trabalhará com os grafismos de um painel, observando sua distribuição em um paredão rupestre. Por fim, a escala de motivo que analisará grafismos de maneira isolada, tendo foco na relação destes com outras representações circundantes ou sobrepostas num mesmo suporte rochoso .

2.4.2 Técnicas de análise e as tradições rupestres

Uma das formas de analisar as gravuras e pinturas rupestres, de modo a facilitar os processos classificatórios, é através do que Pessis define como análise do fenômeno gráfico. A autora, coloca que esta deve ser feita em três dimensões diferentes, são elas: a material; a temática; e a cenográfica.

A primeira destas dimensões é a “material”, a qual trata da técnica por qual cada grafismo foi submetido; a segunda é “a temática”, que como o nome sugere, focará no tema

geral de cada representação, como por exemplo o desenho de uma cena de caça. Por fim, a dimensão “cenográfica”, que terá o foco na observação dos elementos presentes numa cena, a localização destes e a forma como estes se apresentam .

Dentro destas três dimensões, esta pesquisa se aterá a dimensão material, afinal a técnica usada será importante para criar modelagens específicas tanto para figuras gravadas, como para as pintadas. No que se refere às pinturas, outro conceito será usado como *input*, a chamada tradição rupestre.

Podemos definir às tradições rupestres como um agrupamento de diferentes classes de representações que tem características morfológicas e técnicas semelhantes . Estes elementos de análise serão a base para categorizar cada tipo de grafismos em blocos separados de acordo com as suas características. Da necessidade de se criar grupos gráficos que tivessem características culturais semelhantes, independente de quando estes foram produzidos, foram criadas as tradições rupestres (Pessis, 1992).

Martin (2005), destaca duas tradições principais referentes as pinturas, são elas a Nordeste, caracterizada por seus grafismos com cenas bem definidas, com presença de diversos elementos em cada uma delas, retratando de uma forma bem dinâmica o cotidiano do homem pré-histórico; e a tradição Agreste, a qual é representada por cenas estáticas, as vezes com apenas um elemento, marcas de mão, e até representações geométricas.

A Tradição Nordeste, por exemplo, teve sua possível origem a partir de grupos que vieram para o sudeste piauiense, mais especificamente para o PARNA Serra da Capivara, margeando os contornos do Rio Parnaíba. Essas conjecturas referentes a distribuição espacial são limitadas, até por que não há cronologias suficientes, para corroborar este tipo de hipótese .

Ao olharmos esta Tradição mais afundo, tem-se observado que a mesma passou por um processo de diferenciação na sua identidade gráfica ao longo do tempo, o que pode ser corroborado a partir de dados cronológicos indiretos feitos em sítios no PARNA Serra da Capivara, os quais apresentam características gerais semelhantes a Tradição Originária, mas com acréscimos de elementos que fazem com sejam delimitadas como variações da classe, levando ao surgimento de subcategorias ou subtradições .

Essas peculiaridades guardam um padrão gráfico geral que as fez, de início serem elencadas dentro de um grande grupo chamado Tradição Nordeste, com outros desenhos semelhantes encontrados em três áreas distantes do que seria o local de origem deste modo de pintar. É possível encontrar elementos deste agrupamento gráfico no Seridó (RN/PB), no PARNA do Catimbau (PE), e no Vale do São Francisco desde Sergipe até o Norte de Minas Gerais, passando pela área de Central, na Bahia. Os chamados grafismos emblemáticos, como as figuras antropomórficas costa contra costa ou ainda os cabeças de caju, são os melhores exemplos dessas recorrências. .

Além dos grafismos da Tradição Nordeste, há outros grupos bem característicos na Região Nordeste do país, dentre os quais a Tradição Agreste, que também está muito difundida, sobretudo na Região do Pajeú. Esta pode ser caracterizada por seus grafismos puros, desenhos isolados e marcas de mão. Essa última característica inclusive, é um elemento que pode ser encontrado em várias regiões do Nordeste, exceto na Serra da Capivara, o que indica que esta técnica deve ter se originado em outro local, tendo como possível candidato a porção central de Pernambuco. A partir de lá teria se expandido para outros locais, usando sobretudo o vale do São Francisco como uma rota de difusão .

As tradições, porém não podem ser estudadas mais como uma categorização final para qualquer que seja o material, afinal ela acaba por criar uma generalização que pode levar a interpretações errôneas sobre os grupos que produziram aquele determinado vestígio.Van Havre utilizou-se da análise de redes e de correspondências múltiplas para se chegar a marcadores de identidade ou padrões gráficos dos grafismos rupestres analisado.

Matos (2019), por outro lado, buscou criar categorias a partir de análises estatísticas, tais como a análise de componente principal e a cladisítica, observando variáveis específicas em figuras antropomórficas pintadas, tais como largura do traço, movimento, tamanho do desenho, entre outros aspectos. O autor encontrou 10 “padrões gráficos” entre as pinturas do Seridó potiguar, Cariri paraibano e PARNA do Catimbau.

O autor supracitado ainda trabalhou com análises locacionais e identificou relações geoambientais entre os diferentes padrões gráficos e o ambiente em que estavam inseridos. Foi observado que boa parte dos sítios se encontra em regiões de média encosta, tendo alguns sítios com maior ou menor distâncias para drenagens de alto nível e sendo grande parte dos abrigos

em locais de tamanho reduzido, com exceção de um dos padrões gráficos, que tinham áreas abrigadas maiores. Além disso, foi possível perceber que a maioria dos sítios estavam em locais de média a elevada altitude, levando em conta a altimetria das regiões como um todo. Esses aspectos referentes já vem sendo observados no Nordeste Brasileiro com relação aos grafismos rupestres, como é o caso dos trabalhos de Madeira (2016) e Souza (2016).

Esta utilização de recursos estatísticos na Arqueologia e em especial nos estudos do registro rupestre, mostram o potencial deste tipo de análise. O estudo que será aplicado ao Pajeú, terá a intenção de ampliar este leque de possibilidades já existentes e ao mesmo tempo procurar extrair ainda mais informações a respeito dos grafismos rupestres do Nordeste e mais ainda sobre os grupos humanos que por ali passaram.

Além das pinturas, o Pajeú guarda muitos sítios de gravura. Não há ainda uma unanimidade na academia que separem as diferentes gravuras em grupos, sendo a Tradição Itacoatiara, a única tentativa de agrupamento para este tipo de técnica. Esta, porém, foi delimitada com base em sítios como a Pedra do Ingá, que tem representações bem características que não são recorrentes em todos outros sítios gravados .

Os estudos referentes às gravuras rupestres tem indicado uma relação geoambiental bem marcada para estes sítios, os quais estariam diretamente relacionados com cursos d'água , sugerindo uma relação simbólica ou ainda uma necessidade dos grupos pré-históricos em utilizar o recurso hídrico para aplicação da técnica de gravura .

O principal interesse deste trabalho é entender a respeito dos padrões ambientais associados aos sítios arqueológicos. Aqui teremos como referência os locais com registros rupestres, afinal são os exemplares achados em maior número dentro dos limites do Pajeú. Para entender, porém o contexto em que os registros rupestres são encontrados ou como estes se relacionam com o ambiente em que estão inseridos, é necessário pois compreender não só os antecedentes de estudos referentes a estas práticas, como também conhecer os aspectos ambientais, etnológicos e arqueológicos levantados para área de estudo e seus arredores.

3 A BACIA DO PAJEÚ E SEUS CONTEXTOS

3.1 O Contexto etno-histórico

A região da Bacia Hidrográfica do Pajeú, abarca diversos municípios, cobrindo um grande território na porção central do estado de Pernambuco. Esta área, engloba também, sobretudo no sul, diversos agrupamentos de populações indígenas, que perduraram até os dias atuais.

Há uma ideia errônea que coloca os grupos indígenas, sobretudo no nordeste do Brasil, presentes apenas em uma parcela da História, dando uma noção, que após o contato com o colonizador estas populações que aqui já viviam, se misturaram tanto, que acabaram perdendo suas identidades. Isto claramente não é um fato verídico no todo. É evidente que muitos aldeamentos no Brasil colônia tiveram um contato mais intenso com o europeu, porém, muitos nativos, tidos como “rebeldes”, resistiram a ocupação portuguesa em suas terras, e não é à toa, que hoje em dia, muitas etnias possam ser encontradas preservando muito de suas tradições .

É importante contextualizar a situação do nordeste brasileiro, que no início do período colonial, era a região mais importante economicamente da colônia portuguesa. A ocupação inicial se deu nas terras litorâneas e da zona da mata, porém ainda no fim do século XVI, era dado início ao processo de expansão lusitana, usando como rota de deslocamento o Rio São Francisco. Esse processo de interiorização teve logo de ser contido, devido a invasão holandesa a capitania de Pernambuco, período que vai durar até 1654, com a expulsão do flamengos. A partir daí é retomada a política de expansão sertaneja pela Coroa Portuguesa, com o intuito de buscar fontes de metais preciosos e ampliar o povoamento dessas áreas mais afastadas do mar.

Se olharmos para as regiões semiáridas do Nordeste, sobretudo no sertão, muitos grupos indígenas conseguiram resistir aos ataques de invasores, devido sobretudo a dimensão espacial dos sertões, que permitiu que estas populações se dispersassem melhor ao longo do território. É sabido que com a expansão da cana no litoral e zona da mata nordestina, os

europeus se utilizaram, sobretudo do Rio São Francisco, para adentrar ao interior do Brasil e lá se depararam com diversas culturas diferentes .

Com a expansão, o contato com novos grupos de nativos seria inevitável, com isso a própria coroa traça estratégias para lidar com essas populações, primeiramente, usando exploradores bandeirantes, que tinham o intuito de buscar metais preciosos, houve as chamadas "guerras justas", em que era usada a força bruta para conquistar os indígenas, como forma de obter novas mãos de obra nas terras recém exploradas (ARRUTI, 1995; OLIVEIRA; SILVA, 2015).

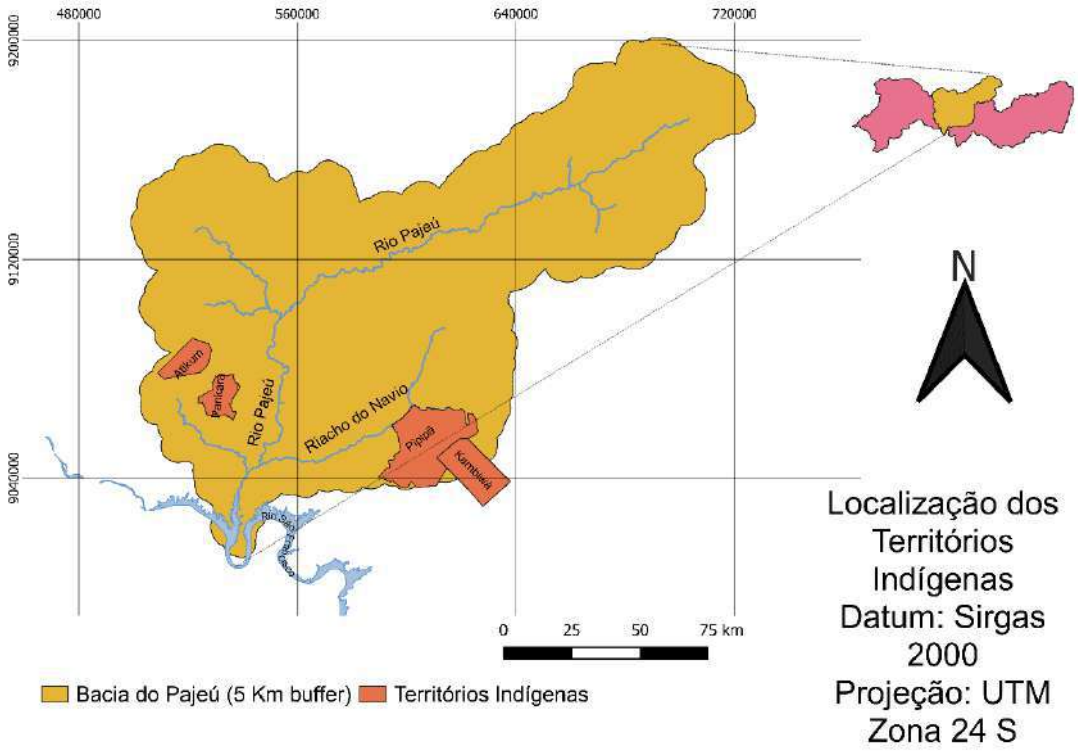
Em período contemporâneo às bandeiras, os missionários católicos, sobretudo jesuítas, adentravam os sertões nordestinos para aplicar as mesmas estratégias já usadas no sul da colônia, na região dos Rios Uruguai, Paraguai e Paraná e no norte, no Maranhão e Amazônia. Eles buscavam reunir indígenas de diferentes etnias da região para trabalhar como mão de obra livre, dentro de uma área específica, sob a supervisão de membros do clero. Essa política também foi incentivada pela Coroa portuguesa, porém diversos embates entre senhores de terra, bandeirantes e missionários foram travados, sobretudo no século XVII. A perda de prestígio dos missionários, sobretudo pela proteção que exerciam sobre os indígenas, acaba por fazer com que muitas missões fossem acabadas ou transformadas em vilas e e os missionários em párocos, com menos poder político (ARRUTI, 1995; PACHECO DE OLIVEIRA; FREIRE, 2006).

Este fato leva a terceira estratégia para extinção das tradições culturais dos grupos sertanejos, caracterizando então um projeto de completa aculturação das populações tradicionais da então colônia portuguesa. A partir de 1755 os casamentos mistos são incentivados e o isolamento indígena que havia durante o período missioneiro, já não mais existe e estes nativos passam por um processo de "nacionalização", é quando se começam a criar a ideia do bom selvagem, tão relatado na cultura romântica do século XIX, largando a vida errante, para se tornar um trabalhador "de fato"(PACHECO DE OLIVEIRA; FREIRE, 2006).

Essa política de extermínio fez com que diversas etnias acabassem por se dispersar pelos sertões, procurando refúgios nas serras, onde membros errantes de grupos distintos inciam um trabalho de cooperação, que no século XX levaria a uma busca pelo reconhecimento legal de seus antigos territórios e tradições. A região do Pajeú, sobretudo no seu baixo curso,

abarca diversos grupos indígenas atualmente, frutos desse processo de mistura, dentro da Bacia e em suas circunvizinhanças, podemos encontrar os Atikum, na Serra do Umã, os Kambiwá e Pipipã, nas Serras Negra e do Periquito e os Pankará, na Serra do Arapuá (**Figura 2**). Esses dados demonstram a grande diversidade do Pajeú em relação a presença de grupos indígenas .

Figura 2: Localização dos territórios indígenas do Pajeú.



Fonte: Elaborado pelo próprio autor (2021)

De modo a entender um pouco da história desses aldeamentos indígenas ainda presentes nos sertões, é preciso fazer um breve recuo cronológico, a partir de 1850, com a promulgação da Lei de Terras, quando boa parte das propriedades rurais começaram a ser regularizadas e passadas para mão de particulares ou tornaram-se áreas devolutas do governo. Esse processo diminuiu ainda mais os já diminutos territórios indígenas no semiárido nordestino, fazendo com que os povos originários “desaparecessem”, enquanto coletividade, no fim do século XIX (OLIVEIRA e SILVA, 2015).

Ao final do século XIX e início do século XX, houve um incentivo para ocupação da região amazônica, sobretudo para ampliação da defesa nacional. Com isso a necessidade na melhora no sistema de comunicação do país, e para isso houve uma comissão, chefiada por militares para instalação de linhas de telégrafos, sobretudo na região Centro-Oeste e Norte do país. Estas equipes de trabalho, se depararam com grupos indígenas, submetidos a uma situação de semi-escravidão, o que levou ao então ao Alferes Cândido Rondon a buscar inserí-los a um trabalho menos exploratório e que os colocassem nas obras de integração nacional aplicadas na região. Com isso, Rondon atuou junto a diversas etnias, usando a mão de obra indígena para ampliar a malha telegráfica do Brasil (PACHECO DE OLIVEIRA; FREIRE, 2006).

Em meados de 1909, deu-se início ao debate sobre a proteção dos grupos indígenas com uma comissão para criação do Serviço de Proteção aos Índios e Localização de Trabalhadores Nacionais (SPILTN), culminando em 1910 com a criação da SPILTN, que viria a ser a SPI. Essa tinha o intuito de transformar a mão de obra indígena em trabalhadores nacionais, tirando a Igreja das aldeias e realizando um outro tipo de abordagem, que visava, de início proteger esses índios, cedendo um território para realização de suas atividades (PACHECO DE OLIVEIRA; FREIRE, 2006).

Nesse contexto em que diversas etnias Brasil a fora tinham seus pedaços de terra sob guarda da União garantido por lei é que em 1920, os Fulni-ô, que habitam hoje o agreste pernambucano, tiveram seu aldeamento reconhecido pelo Serviço de Proteção do Índio – SPI, pois, segundo os avaliadores da época, serem um dos únicos grupos do Nordeste a ter preservado sua língua e serem vistos como “os mais claramente índios entre as populações indígenas do Nordeste” (SILVA, 2017a).

Após o evento descrito acima, muitos outros grupos passaram a reivindicar seus territórios, e as populações que habitavam a Bacia do Pajeú não foram diferentes. Houve uma imensa cooperação entre os diversos grupos indígenas da região para garantir o reconhecimento das diferentes etnias ali localizadas, os Atikum, por exemplo, lutam desde 1930 pelo reconhecimento de suas terras e tiveram que realizar um tipo de união com os grupos que habitavam a Serra do Arapuá, atualmente ocupada pelo Pankarás, para conseguir este feito (SILVA, 2017b).

Os líderes Atikum, resolveram iniciar o processo de regularização do seu território e fizeram, junto a SPI, a solicitação de uma visita à região, porém, o grande sinal diacrítico de indianidade, segundo Silva (2017a), era o ritual da toré, todavia, nem todos os integrantes Atikum dominavam a arte, por isso nativos da Serra do Arapuá e os Tuxá de Rodelas, na porção baiana do São Francisco, foram chamados para compor o grupo. Ao final, os Atikum, tiveram suas terras demarcadas em 1949 e passaram a ter relações amicais com os grupos da Serra do Arapuá e de Rodelas.

No fim do século XX, os grupos da Serra do Arapuá, iniciam um processo de resgate identitário, que culmina, em 2003, no reconhecimento, por parte dos habitantes da Serra do Arapuá, que passariam a ser denominados agora de Pankarás. Estes, porém ainda não conseguiram a demarcação oficial de suas terras.

Os Pankarás, e outras etnias do Pajeú, mesmo os que têm suas terras demarcadas, sofrem com a constante pressão do agronegócio e das obras de infraestrutura que ocorrem no sertão. Um grande exemplo dessa resistência indígena, é com relação as obras da transposição do Rio São Francisco.

Camargo , em sua tese, narra sua visita ao município de Floresta, durante as obras de implantação do Canal Leste e da Adutora do Pajeú na região. A autora relata a situação tensa entre os trabalhadores rurais do município, sejam de origem indígena, quilombola e assentados da reforma agrária, que lutam cada qual por seu pedaço de chão. Com o início das obras de transposição do São Francisco, alguns territórios, ou áreas habitadas por indígenas foram atingidas, como foi o caso dos territórios Atikum.

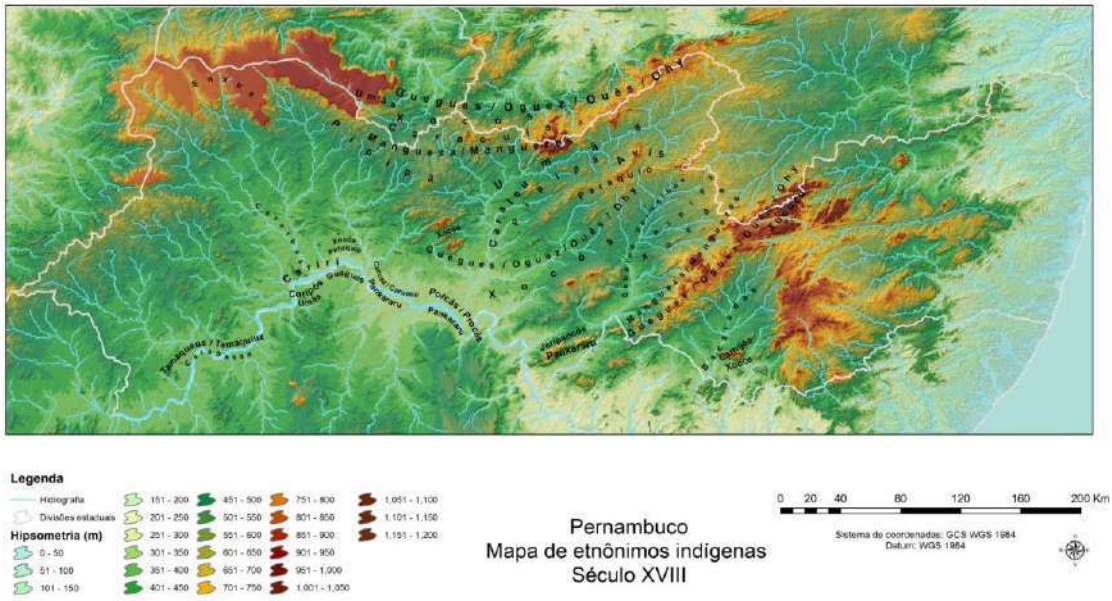
Toda essa questão histórica, é importante ser compreendida, afinal, mostra toda uma dinâmica no espaço de ocupação humana na região do Pajeú. É preciso, também, dar voz a população indígena, primeiros habitantes destas terras, e que tem um papel importante, na preservação das tradições culturais e da história local.

A dinâmica espacial dos grupos indígenas também pode ser vista através da análise de dados etnográficos com o foco em compreender etnias que viveram no passado, sendo algumas não mais citadas como grupos que perduraram até os dias atuais. Medeiros e Mutzenberg fizeram um apanhado de fontes documentais primárias e secundárias de cronistas,

e conseguiram realizar um mapeamento dos grupos indígenas que habitaram o estado de Pernambuco no século XVIII.

Os autores supracitados conseguiram desenvolver um mapa com diversos grupos, divididos cada uma dentro das unidades ambientais que ocupavam a época (**Figura 3**). Se olharmos para a região da Bacia do Pajeú, identificamos os Avis, Caracuis, Carateús, Guegues, Manguezes, Paraquiós, Pipipãs, Umãs e Xocós. Muitos desses grupos são registrados em diversas áreas que vão ao norte da Bacia do Pajeú, na Serra dos Cariris Velhos, passando por áreas ao longo dos Rios Pajeú e Moxotó e também ao largo do submédio São Francisco e nas ilhas que este rio forma.

Figura 3: Mapa de distribuição dos grupos indígenas do Estado de Pernambuco no século XVIII



Fonte: MEDEIROS E MUTZENBERG, 2013

O sertão do Pajeú também é retratado por Pereira da Costa nos Anais Pernambucanos , como sendo um território dos Rodelas, grupo este, que segundo Silva (2003), habitava a porção baiana do São Francisco, em frente a foz do Pajeú e pode ter sido confundido com os Avis, que habitavam o lado pernambucano. Pereira da Costa, porém, fala a respeito do intenso deslocamento de grupos indígenas, que migravam em busca de terras mais úmidas e

mais férteis ao longo das margens do Pajeú, corroborando o intenso fluxo migratório que ocorria na região.

3.2 Contexto arqueológico no Pajeú

Além dos estudos etno-históricos, alguns estudos arqueológicos foram realizados na Bacia do Pajeú e estão relacionados, primeiramente, a obras de infraestrutura do setor elétrico ocorridas ao longo do médio curso do Rio São Francisco. Ainda no ano de 1913, o empresário alagoano Delmiro Gouveia, construiu a primeira hidrelétrica do Rio São Francisco, Angiquinho, esta, porém, não está mais em atividade, servindo hoje apenas como museu .

Em 1945, através de um decreto do Presidente Getúlio Vargas, foi proposta a criação da CHESF – Companhia Hidrelétrica do São Francisco e com ela a missão de dar continuidade a obra inicial de Delmiro Gouveia para aproveitar o potencial do Rio São Francisco em produzir energia. Com isso, no ano de 1955, a inauguração da Usina Hidrelétrica de Paulo Afonso I, marca o primeiro projeto da CHESF, ao longo do Rio São Francisco .

A obra de maior impacto para a bacia do Pajeú, sobretudo para seu baixo curso e foz, foi a construção da Usina de Itaparica, pois o lago do reservatório compreende uma área que afeta diretamente a chamada Barra do Pajeú. Essa obra fez com que boa parte das margens do São Francisco fossem inundadas e diversos sítios arqueológicos ficaram submersos .

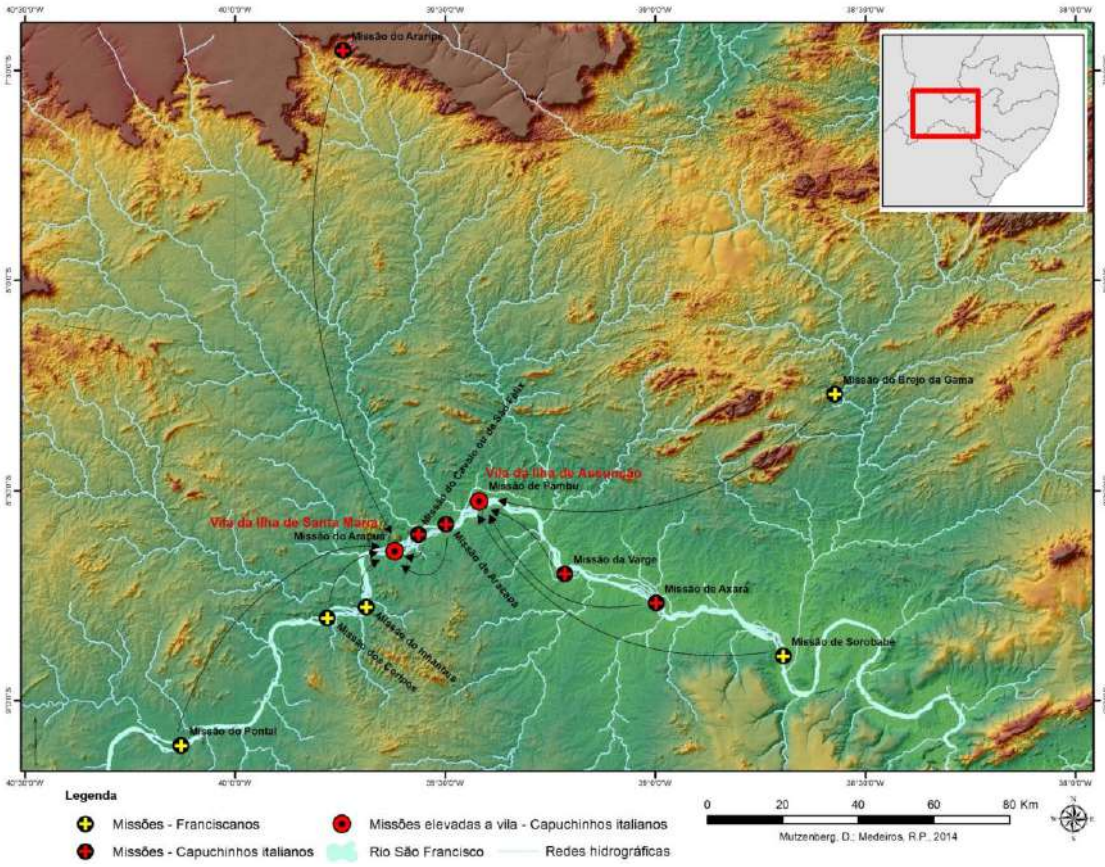
Para realização do resgate desses remanescentes arqueológicos que seriam impactados pela construção da usina, foi realizado entre os anos de 1981 e 1988, diversas campanhas nas margens pernambucanas e baianas do São Francisco, a fim de encontrar e salvar tudo o que fosse possível a respeito da presença do homem na região.

Durante o período de trabalhos na região foram encontrados diversos sítios com gravuras rupestres ao longo das margens, sobretudo nas proximidades do município de Petrolândia, 50 km a leste da Foz do Rio do Pajeú. Segundo Martin , foram evidenciados dois tipos de assentamentos de caçadores coletores na região, uns localizados em áreas abertas com presença de material lítico, mas uma estratigrafia pouco profunda, aparentando serem locais de breves ocupações. Por outro lado, os abrigos sob rocha, apresentavam uma longa seção estratigráfica, o que os caracterizava como locais de ocupações mais persistentes.

Outros locais de ocupações mais duradouras foram as ilhas formadas pela deposição sedimentar na porção do submédio São Francisco, apontados por Martin (1998) e Etchevarne como um local utilizado pelos grupos para um assentamento longo, onde eles se aproveitavam da fertilidade dos solos para o plantio de diversas culturas. Nessas porções de terras, também foi encontrado além dos vestígios líticos, artefatos cerâmicos, que segundo Etchevarne (2002) permitiu estabelecer uma semelhança cultural com os materiais encontrados nos sítios das margens do São Francisco.

As ilhas do submédio São Francisco foram, quase que em sua totalidade, cobertas pelas águas dos lagos das usinas hidrelétricas e por isso todo o material que se tem hoje sobre elas foram fruto das pesquisas que ocorreram na década de 1980. Medeiros e Mutzenberg (2014) fizeram o mapeamento de todas essas ilhas, através de um estudo de cartografia histórica para perceber a espacialidade dos grupos indígenas e das missões catequizadoras que se instalaram sob esses locais após a chegada dos europeus (**Figura 4**).

Figura 4: Mapa de localização das missões jesuíticas do submédio São Francisco.



Fonte: MEDEIROS E MUTZENBERG, 2014

Neste trabalho descrito acima podemos destacar dois aldeamentos — que posteriormente viriam a ser missões catequizadoras — ao longo da Bacia do Pajeú, um nas proximidades da atual Serra do Arapuá, a Missão do Brejo da Gama, nas margens do Rio Pajeú; outro na Ilha de Sorobabel, bem na Barra do Pajeú — que posteriormente se tornou a Missão de Sorobabé. Nas duas localidades, as ações da Igreja foram coordenadas por freis franciscanos, sendo que na primeira, o aldeamento anterior a chegada dos europeus era formado por indígenas Pipipãs e na segunda por indígenas Porcos e Brankararus.

A Ilha de Sorobabel é um dos locais mais estudados dentre as ilhas do submédio São Francisco. Esta ilha que era uma das maiores em extensão no Rio, tem dois sítios arqueológicos, um mais relacionado a fase pré-contato e outro já com elementos que indicavam a mistura de culturas, após a instalação da Missão de Nossa Senhora do Ó (Missão de Sorobabé)

.

O sítio pré-contato, conhecido como Ilha de Sorobabel, é caracterizado como sendo um aldeamento indígena com a presença de cerâmicas de engobo vermelho dispersos em uma circunferência de 50m ao redor de uma mancha circular (SILVA, 2003). Segundo Martin (1998), as escavações na Ilha se deram após o relato de moradores nas regiões que retrataram a presença desses materiais cerâmicos. No local chegou a ser encontrado algumas urnas funerárias com restos humanos além de colares de contas, pingentes e outros adornos.

No sítio Nossa Senhora do Ó, porém, os vestígios pré-históricos se confundem com os de período histórico. Lima, Medeiros e Castro (2016) observaram que houve uma continuidade no perfil técnico entre as cerâmicas encontradas no sítio Ilha de Sorobabel e no Missão Nossa Senhora do Ó, indicando que mesmo com as influências europeias, os indígenas devem ter continuado sua tradição.

Ao olharmos o Cadastro Nacional de Sítios Arqueológicos (CNSA/IPHAN) até 2018 , apenas 8 sítios estavam cadastrados na base e tinham pontos de geolocalização, todos estes com presença de grafismos rupestres. Na planilha referente aos dados não georreferenciados, também são poucos os sítios encontrados, sendo apenas 7 registros para a Bacia do Rio Pajeú, e a maioria deles relacionados ao registro rupestre.

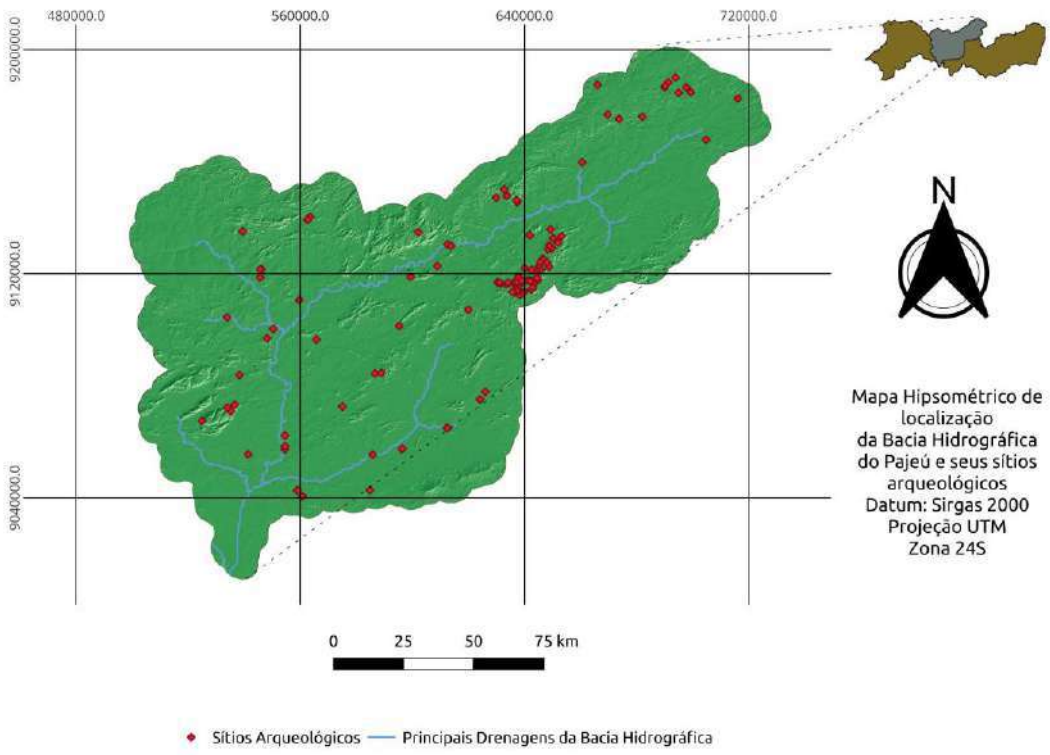
Como é possível ver, a região do Pajeú conta com poucas pesquisas sistemáticas de Arqueologia e o registro de sítios na base de dados do IPHAN é bastante escasso. Afim de diminuir essas lacunas, algumas pesquisas se voltaram para a Bacia, como é o caso de Barbosa, Cisneiros e Mutzenberg , por exemplo, trabalharam com as representações antropomórficas em um sítio da Bacia do Rio Pajeú, porém o enfoque do estudo eram os grafismos do Vale do Moxotó. O projeto de Pessis et al (2017), que é a maior fonte de dados desta tese, também tem uma importância para ampliação acerca do conhecimento arqueológico na área.

Uma última menção aos grafismos rupestres do Pajeú foi feita em um parecer técnico , em que os sítios do município de Flores, mais especificamente os localizados ao longo do Riacho das Letras, foram fotografados e descritos, mas sem nenhum tipo de análise mais aprofundada.

É interessante perceber que durante os trabalhos relatados por Pessis et al (2017), 155 sítios com grafismos foram levantados dentro dos limites da Bacia do Pajeú e áreas

imediatamente adjacentes (**Figura 5**). Neste estudo, olhando para o alto curso do Rio Pajeú, pode-se observar uma área de grande concentração de sítios, na margem esquerda do Rio, sobretudo, na chamada Serra Vermelha.

Figura 5: Mapa Hipsométrico de Localização da Bacia Hidrográfica do Pajeú e seus sítios arqueológicos.



Fonte: Elaborado pelo próprio autor (2022)

Ao passo que o Pajeú tem poucos trabalhos relatados na literatura, outras áreas imediatamente adjacentes a ele contam com uma vasta gama de teses, dissertações e artigos referentes aos vestígios arqueológicos lá encontrados. As regiões do Moxotó e do Ipanema, a leste do Pajeú, por exemplo, são locais com intensas pesquisas arqueológicas desde a década de 80 e que trazem uma cronologia da ocupação da área.

Os sítios Mingú e Quixéu I e II, localizados dentro dos limites do PARNA do Catimbau, na Bacia do Ipanema, apresentaram cronologias de cerca de 6640 ± 95 a 2780 ± 190 AP, referentes aos enterramentos primários presentes nos abrigos sob rocha . Ainda no PARNA

do Catimbau outros dois sítios também ofereceram datações a partir de seus remanescentes, é o caso do sítio Pedra do Cachorro , com datas que variam de 2100 ± 30 a 640 ± 30 AP e o sítio Alcobaça que apresenta um longo período de ocupação baseada na datação por radiocarbono, variando entre 4851 ± 30 e 888 ± 25 AP . Vale ressaltar também que este sítio contém um grande painel rupestre com pinturas características da tradição agreste, o que pode sugerir a antiguidade deste modo de fazer pintura.

Do outro lado do Pajeú, a oeste, na região da Bacia do Rio Terra Nova, cronologias também são conhecidas, mas referentes a sítios arqueológicos em contextos diferentes. O município de Salgueiro guarda dois sítios a céu aberto importantes para entender o contexto dos grupos que habitavam o semiárido pernambucano. No distrito de Conceição das Creoulas, na Lagoa da Pedra, foram encontrados vestígios líticos associados a ossos de megafauna, alguns deles com marcas de descarne.

Já bem próximo à sede do município de Salgueiro, às margens do canal da transposição do Rio São Francisco, foi encontrado um outro sítio, com presença de megafauna e vestígios humanos, porém não há estudos que relacionem os achados paleontológicos e arqueológicos, mas há sim, cronologias obtidas por LOE, que variam para as camadas onde foram encontrados materiais líticos, entre 13.500 e 11.000 AP. Essas datas trazem a ocupação do homem naquela região do atual nordeste brasileiro bem para trás no tempo .

É importante reforçar, porém, que os sítios do Pajeú em si não têm nenhum tipo de cronologia associada a eles e aqui faremos um “achatamento” analítico, de modo a colocar todos os locais com grafismos rupestres dentro de um mesmo grupo temporal, a única distinção será observada nas análises entre pinturas e gravuras, como anteriormente mencionadas.

3.2.1 Os sítios rupestres da região

A região do Pajeú conta com sítios de grafismos rupestres bastante diversos que vão desde gravuras com formas geométricas e grafismos puros, sítios com pinturas rupestres de marcas de mão e também de figuras antropomórficas isoladas. Alguns exemplos desses desenhos podem ser vistos nas imagens abaixo (**Figuras 6, 7, 8 e 9)**. Temos respectivamente os sítios de gravura Pedra do Riacho da Velha, no alto curso do Rio Pajeú, próximo a sede do município de Flores e o Mãe d’Agua I, no baixo curso, próximo a sede do município de Floresta.

Figura 6: Painel com Gravuras Rupestres do sítio Pedra do Riacho da Velha, no município de Flores.

Foto tirada pelo próprio autor (2022)

Figura 7: Gravura rupestre do sítio arqueológico Mãe D'água I, no município de Floresta.

Foto tirada pelo próprio autor (2021)

Os outros dois, em sequência, são a Pedra do Catolé e o Pedra da Cuia, ambos no município de Carnaubeira da Penha, nas próximidades do território Pankará.

Figura 8: Marcas de mão pintadas visualizadas no sítio Pedra da Cuia, no município de Carnaubeira da Penha.

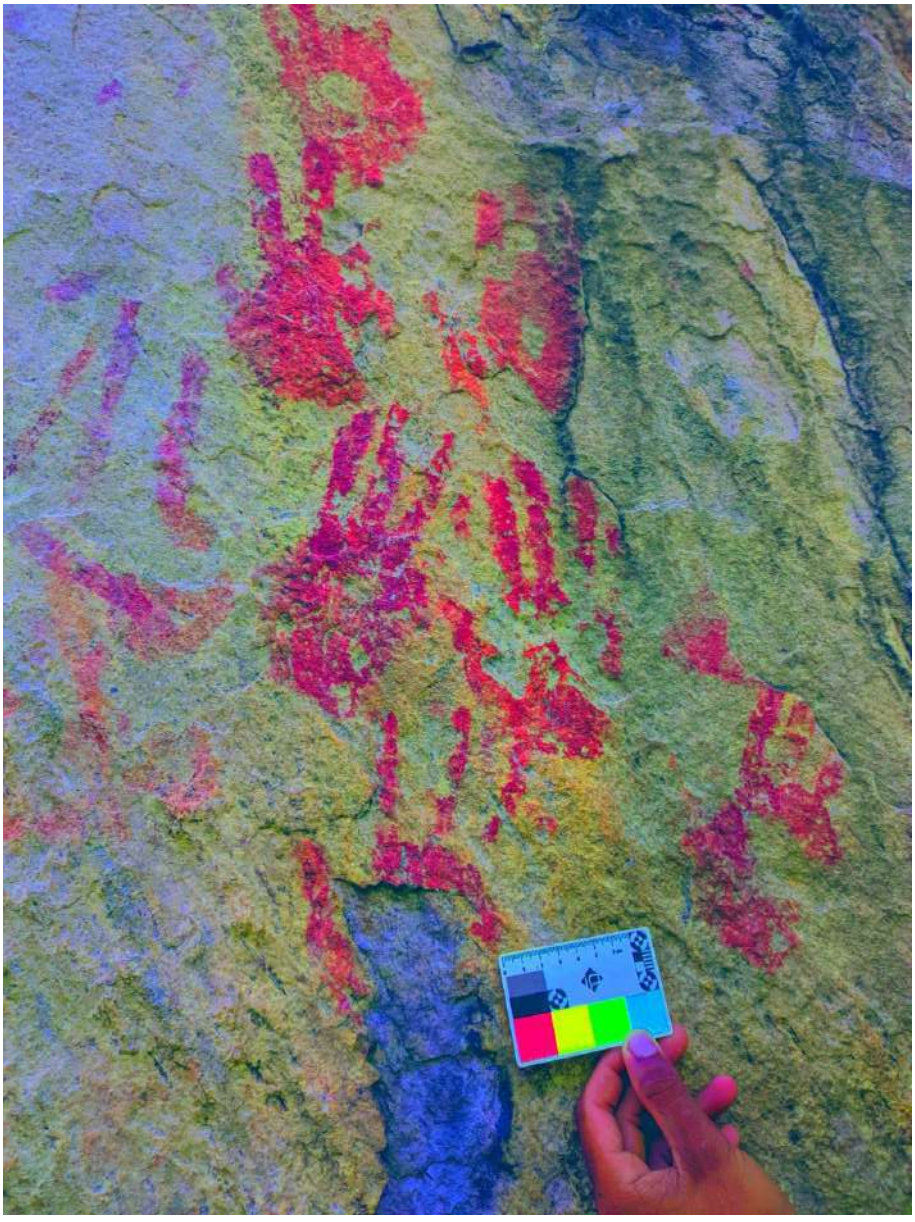


Foto tirada pelo próprio autor e tratada no software Rock Art Enhancer (2021)

Figura 9: : Foto de grafismos puros pintados no sítio Pedra do Catolé, no município de Carnaubeira da Penha.

Foto tirada pelo próprio autor e tratada no software Rock Art Enhancer (2021)

3.3 Contexto ambiental

Este trabalho envolverá uma área de estudo que engloba toda a Bacia do Pajeú. Esta é localizada inteiramente na região semiárida, de acordo com Superintendência de Desenvolvimento do Nordeste (SUDENE), que leva em conta características climáticas de semiaridez e baixa precipitação pluviométrica. A hidrografia regional também é composta em sua maioria de rios intermitentes e efêmeros, que devido ao baixo ciclo pluviométrico não sustentam uma grande quantidade de água, com exceção ao Rio São Francisco, que por ser, muitas vezes, a única fonte de água no local, ganha uma importância maior para as populações sertanejas .

A porção semiárida abarca uma vasta extensão do território nordestino, abrangendo todos os estados da região, com exceção do Maranhão, e inclui ainda, a parte Norte do estado de Minas Gerais. A população local convive anualmente com períodos de grande seca, o que torna a vida desse povo um desafio diário, desde tempos remotos, anteriores mesmo a colonização .

Voltando os olhares para a Bacia do Pajeú, esta abrange um território de aproximadamente 16.600 km², perpassando um total de 26 municípios, dentro do quadrante 07°16’20” e 08°56’01” de latitude sul, e 36°59’00” e 38°57’45” de longitude oeste. A nascente do Rio Pajeú está localizada na Serra do Balanço, no município de Brejinho, a cerca de 800 m de altitude. De lá, o rio percorre 353 km, até sua foz, nas proximidades do Lago da Hidrelétrica de Itaparica, no Rio São Francisco .

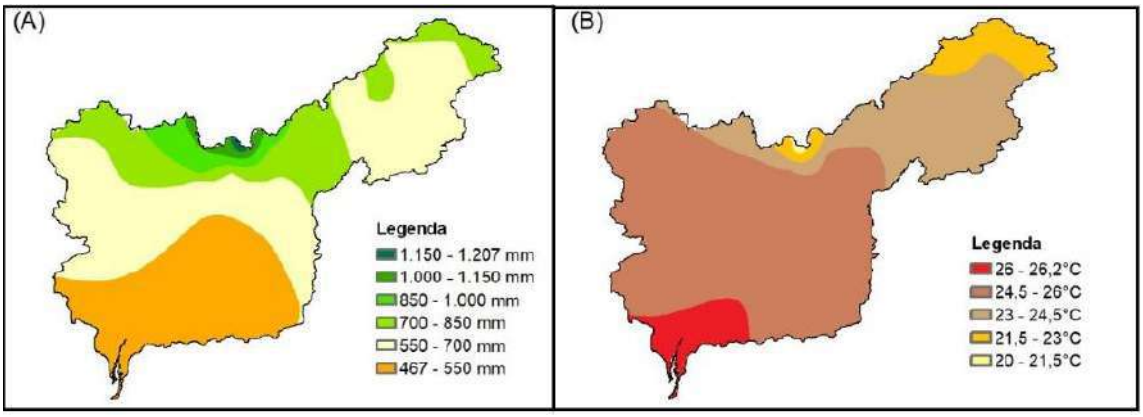
O clima predominante na região é o semiárido quente e seco com volumes pluviométricos que variam entre 550 a 800 mm (**Figura 10**), porém nos pontos mais altos, como a área da Serra da Baixa Verde, que envolve os municípios de Triunfo e Santa Cruz da Baixa Verde, às chuvas podem aparecer com mais intensidade, ultrapassando os 1000 mm .

O controle climático, e consequentemente pluviométrico, da região semiárida do Nordeste como um todo, está atrelado sobretudo a três fatores. o principal entre estes é a ocorrência da Zona de Convergência Inter Tropical (ZCIT). Esse sistema atua no Oceano Atlântico e é regulado pelas variações de temperaturas deste. Quando a porção norte equatorial do Atlântico está quente, ela impede a formação de chuvas na porção Norte da Região Nordeste,

que é a mais árida. A medida que o mar esfria a norte, a ZCIT se move para sul, provocando chuvas na parte setentrional do semiárido nordestino .

Os outros dois fatores descritos por Corrêa et al (2019) para dinâmica climática da região semiárida do Brasil são: O *La Niña*, que resulta no esfriamento das águas do Oceano Pacífico e influencia na formação de nuvens de chuva no Nordeste brasileiro e a Topografia. Em determinados pontos do semiárido há ocorrências de maciços residuais que se destacam na paisagem, com cotas superiores ao entorno, o que provoca a ocorrência de chuvas orográficas, nas porções a barlavento da encosta.

Figura 10: Precipitações médias e temperaturas médias do Pajeú. A) Volume precipitação anual. B) Temperatura média anual.



Fonte: (SOARES; NÓBREGA; GALVÍNCIO, 2018)

A vegetação predominante, como em todo o semiárido nordestino, é formada de caatinga, tanto hiperxerófila como floresta caducifólia . Nas porções mais elevadas do terreno, a vegetação pode ser caracterizada como Floresta sub-caducifólia, com resquícios de plantas típicas de mata atlântica (TAVARES, 2015).

O relevo do Pajeú apresenta uma grande amplitude de cotas altimétricas, na porção mais elevada, atinge cerca de 1200 m acima do nível do mar, no município de Triunfo, mais precisamente no Pico do Papagaio, que é também um dos pontos mais elevados do Estado de

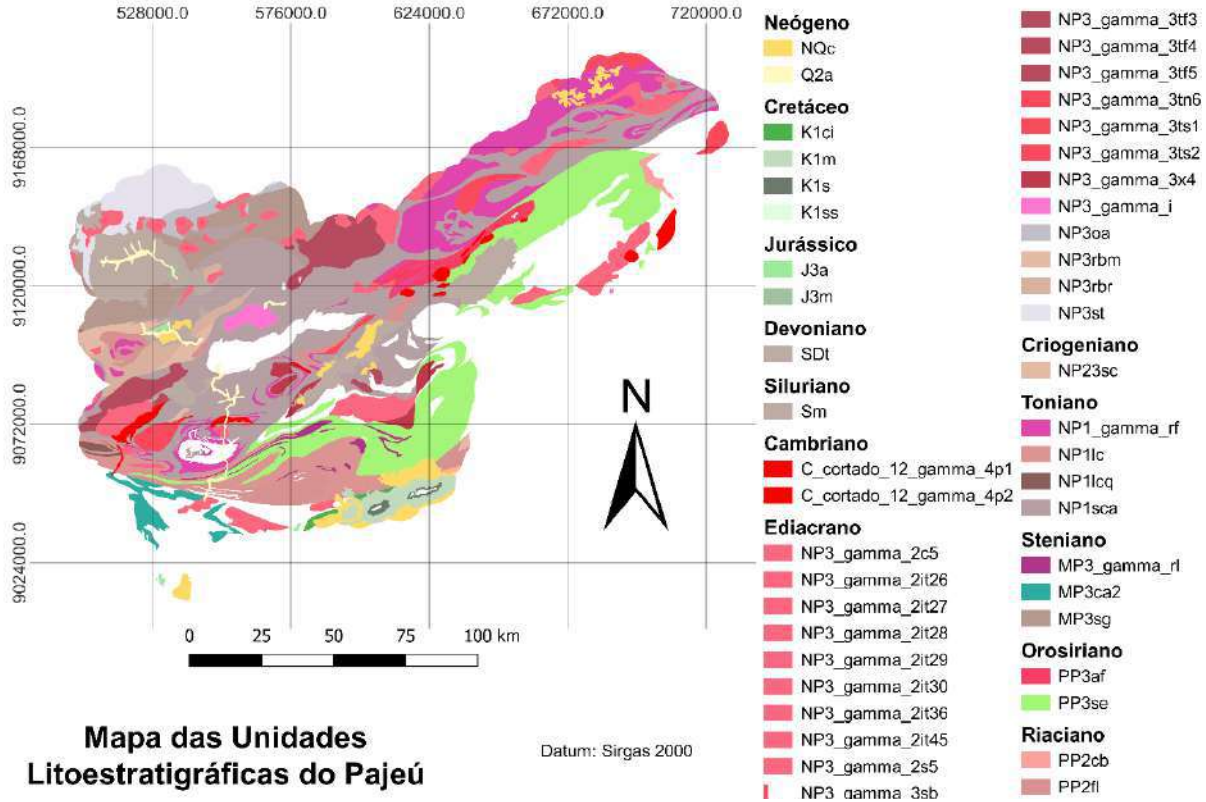
Pernambuco. Já nas porções mais baixas, na Depressão São Franciscana, as cotas podem cair para cerca de 250 m (SILVA, 2019).

Em relação a Geomorfologia, a Bacia está inserida em duas unidades geoambientais. O alto curso do Rio está inserido no Domínio Geomorfológico do Planalto da Borborema, caracterizado pela presença de relevos montanhosos e ondulados (RIBEIRO, 2014). Esta região tem por característica apresentar áreas elevadas que variam até 500m acima das cercanias mais baixas e sua origem remonta a quebra do continente Gondwana e posterior magmatismo entre placas, do Mioceno, que reativou estruturas herdadas (CORRÊA et al, 2019).

O baixo curso do Rio Pajeú está inserido numa porção de terras rebaixadas, de cotas altimétricas bem inferiores aos do alto curso, associadas a estruturas metamórficas e metassedimentares, com presença de enclaves Meso e Paleoproterozóico que favorecem a presença de inselbergues e maciços residuais (intrusões neoproterozoicas concebidas pelo intenso processo erosivo e alterações climáticas) (CORRÊA et al, 2019) tidas na literatura como locais de refúgio de grupos humanos e da biodiversidade .

Com relação a litologia, a bacia do Pajeú é composta sobretudo de rochas metamórficas, tais como os meta granitos, metagnaisses, mármores, metabasaltos, entre outras (**Figura 11**). Em algumas áreas, porém, há ocorrências de rocha sedimentar (CPRM, 2017), inclusive, nesses locais, a presença de registros rupestres é a mais frequente de toda a Bacia.

Figura 11: Litoestratigrafia do Pajeú



Fonte: Elaborado pelo próprio autor (2022)

3.4 Contexto paleoambiental

O estudo da evolução das paisagens geomorfológicas foi iniciado nas ciências geográficas por meio das pesquisas de William Morris Davis, com sua obra, “O Ciclo Geográfico” (1889). Esse trabalho foi o pontapé inicial para que muitos autores passassem a dar valor a história do ambiente. Davis recebeu muitas críticas por levar a variável tempo muito a fundo e esquecer de características fundamentais como as estruturas geológicas e os processos geomofológicos. Em síntese, Corrêa (2001), coloca que a Geomorfologia busca entender a causa pelo efeito, colocando sedimentos correlativos como a consequência para revelar a história do ambiente.

No Brasil, mais especificamente na região Nordeste, os estudos para compreensão dos processos geomorfológicos são bem destacados, sobretudo na região do Planalto da Borborema, local que abarca boa parte área de estudo. Estas pesquisas têm por muitas vezes o Período Quaternário como foco, que segundo Suguio (2010), é o período geológico vigente e inicia-se por volta de 2,5 milhões AP. Ele é caracterizado por apresentar depósitos sedimentares com restos de animais e vegetais, que na grande maioria são encontrados ainda hoje. Esse período é subdividido ainda em Pleistoceno e Holoceno, sendo a segunda parte em que se concentra a maior gama de estudos arqueológicos na América.

Para se entender o paleoambiente é preciso compreender primeiramente como era o clima do passado, pois o mesmo influencia diretamente na formação e modificação do relevo. Como é trazido por Galvão , os eventos glaciais e interglaciais resultam na oscilação glácio-eustática do nível do mar e mudanças ambientais.

As pesquisas paleoambientais para o Nordeste do Brasil vem sendo ampliadas graças sobretudo ao uso das diversas técnicas de análise hoje disponíveis. Behling et al , chegaram a conclusões importante a respeito do paleoclima no Nordeste Brasileiro. Graças a dados de isótopos de oxigênio, e palinológicos, pôde-se perceber que o Nordeste, desde 40 mil anos AP, passou em sua maioria por períodos de seca, com alguns episódios de maior umidade, há 40.000, 33.000, 24.000 e entre 15.500 e 11.800 AP.

Dentro das pesquisas de reconstruções paleoclimáticas no Nordeste do Brasil, um tipo de feição geomorfológica vem se tornando bastante importante no estudo de reconstruções

paleoambientais, são as chamadas “marmitas de dissolução”, que são depressões escavadas na rocha fresca. Essas estão associadas a rochas com baixa coesão e resistividade às intempéries físicas e químicas .

No semiárido a ocorrência de marmitas está relacionada às áreas de maior umidade, como Brejos de Altitude, porém é possível encontrar um certo número de depressões associadas ao embasamento cristalino exposto, que servem como área de deposição sedimentar produzida por sistemas de drenagens inativos nos dias atuais, tendo, portanto, características semelhantes as marmitas (SILVA, 2013).

Um dos exemplos mais recentes para compreensão desse paleoambiente está na Lagoa Uri de Cima, no município de Salgueiro, área bem próxima a Bacia do Pajeú. Em dados recente colhidos no local, foi possível perceber que houve eventos de grande magnitude pluviométrica, por volta 12 mil anos AP , o que pode ser observado na seção abaixo (**Figura 12**). Após isso o clima na região muda drasticamente, se tornando mais árido. Ao longo dos últimos 10 mil AP já mais próximos do clima atual, o que explica a formação de uma camada de concreção carbonática — que deve ter se formado a cerca de 9 mil AP — na porção logo abaixo a esta seção mais escura, associada ao evento mais úmido .

Figura 12:Seção estratigráfica Lagoa Uri de Cima

Fonte: MÜTZENBERG et al., 2013

Esse período mais úmido, mostrado na figura acima, com datações entre 13 e 11 mil pode ser corroborado por uma grande presença de vestígios ósseos de cágados, animais que precisam de um ambiente de maior umidade, junta-se a isso, o achado de artefatos líticos nestas camadas, indicando uma presença humana .

Ao olharmos para os períodos mais recentes, no Holoceno tardio (2500 AP), Utida et al , em um estudo de uma lagoa na cidade Camocim, no Ceará, percebem que houve alternâncias entre climas secos e de maior umidade. Através da análise de isótopos de hidrogênio, chegou-se à conclusão de que de 2500 a 2420 AP havia um período de maior umidade no Nordeste brasileiro. Entre 2000 AP e 700 AP, houve um novo processo de

aridificação. Por fim, já em tempos históricos, entre 500 e 100 AP, o clima fica um pouco mais úmido, sendo observadas taxas próximas as observadas no primeiro conjunto de datas.

Uma das grandes características observadas na maioria dos trabalhos a respeito do paleoclima do Nordeste é a associação e a observação na variação da localização da ZCIT ao longo do atlântico. São observados dentro dessas análises, períodos em que este sistema meteorológico permanece um bom tempo bem ao norte da América do Sul e nos períodos mais chuvosos do Semiárido, ela desce, chegando até a costa da Bahia (BEHLING et al, 2000; UTIDA et al, 2019).

Para ter um panorama geral do paleoambiente nordestino, Missura , faz um apanhado geral de 37 trabalhos que enfocam neste tema. Todos esses, têm como base, datações absolutas e resultados diversos, que mostraram, no geral, números em comum, que colocam o Pleistoceno Tardio, em específico o período do Último Máximo Glacial (20 mil AP); a fase de alternância Pleistoceno/Holoceno (entre 12 e 10 mil AP); e o Holoceno Médio como um período de clima mais seco, com diminuições significativas na intensidade das chuvas, as quais não tinham capacidade de remobilizar os sedimentos profundos, até a fase de alternância com o Holoceno (20 à 10 mil AP), que se caracteriza como uma época mais úmida, até meados do Holoceno Médio quando as vegetações mais densas começam a dar espaço a uma vegetação mais próxima ao cerrado e a caatinga e no Holoceno Superior é estabelecida a vegetação atual.

Observar esses elementos que compuseram ou compõe a paisagem terrestre é importante em qualquer trabalho de Arqueologia, nesta pesquisa, esses dados terão um peso ainda maior, pois eles serão norteadores para o desenvolvimento e criação do modelo preditivo e permitirão extrair as informações espaciais necessárias para entender a distribuição dos sítios arqueológicos com grafismos rupestres na Bacia do Pajeú.

Com os dados de pesquisas acerca do paleoambiente do Nordeste, fica claro que os grupos pré-históricos, que fizeram os grafismos rupestres da Região, conviviam constantemente com períodos de estiagem, o que deve ter feito com estes se estabelecessem próximo das principais drenagens da Bacia. Observar essa situação em contraste com a localização dos sítios arqueológicos, pode ser importante para melhor entender a dinâmica dos grupos pretéritos, mesmo que não haja cronologias associadas a estas representações.

4 MÉTODOS

Como já foi tratado nos capítulos anteriores, as pesquisas que trabalham com o registro rupestre, tem buscado novas metodologias para se chegar a respostas mais concretas a respeito desse tipo de vestígio arqueológico. Este estudo, por sua vez, tratará de elaborar uma análise espacial, a princípio em uma escala macro, envolvendo os sítios arqueológicos com pinturas e gravuras de toda a Bacia do Pajeú, e para isso serão utilizadas ferramentas, tais como: os SIG, dados de sensoriamento remoto e algoritmos de AM.

É importante ressaltar, porém, que aqui trabalharemos com pinturas e gravuras colocados dentro de uma mesmo conjunto cronológico e técnico (a princípio), já que não temos nenhuma informação de datação a respeito dos sítios do Pajeú — apesar de algumas datas associadas a sítios de registros ou de outras materialidades em locais vizinhos serem conhecidas — mas que de certa forma apontam para períodos diferentes de confecção de determinados vestígios, porém aqui serão tratados como um grupo único.

No que tange a técnica empregada, sejam referentes aos diferentes estilos de se pintar ou ainda o uso de registros gravados, estes também são tratados como pertencentes a um mesmo grupo, ao menos incialmente, já que temos uma base de dados muito escassa e separá-los de início só gera perda de informação. Isso não nos impediu, porém, de executar modelos separados para as duas técnicas afim de averiguar possíveis diferenças que possam ser encontradas, já que este é um tema que vem sendo debatido em outras pesquisas .

O levantamento bibliográfico para o desenvolvimento metodológico desta pesquisa teve como base a literatura da Tecnologia da Informação, já que na Arqueologia como um todo, as aplicações que se utilizam de algoritmos de AM, começaram a se desenvolver na última década. Ainda assim, a leitura de diversas utilizações no cenário arqueológico foi de grande auxílio no proceder com os dados aplicados nesta tese.

Primeiramente, vale aqui reforçar, que os dados usados nesta pesquisa partem da ideia da dicotomia presença/ausência, ou seja, aqui trabalharemos tanto com os dados que representam os sítios arqueológicos obtidos através do Projeto de caracterização e documentação dos sítios rupestres no semiárido de Pernambuco como também a ausência destes , partindo dos dados coletados nas atividades de prospecção.

Neste capítulo passaremos pelas diversas fases da construção de um Sistema de Informação Geográfica para os sítios com grafismos rupestres presentes na Bacia do Pajeú e na elaboração do MPA para a região. Partiremos desde a obtenção e tratamento dos dados, até a fase de teste da confiabilidade do modelo obtido após as visitas a campo.

4.1 Um pouco mais sobre o SIG na Arqueologia

Como já foi comentado anteriormente, os SIG são uma ferramenta de suma importância na compreensão dos contextos arqueológicos e nas análises espaciais. Segundo Gallotti et al. são o maior exemplo da evolução tecnológica no que diz respeito à análise espacial em Arqueologia, pois os mesmos contêm ferramentas que facilitam o gerenciamento dos dados arqueológicos e aparecem como uma peça importante no entendimento dos contextos arqueológicos.

Conceitualmente, os SIG são sistemas computacionais usados para o armazenamento, gestão, recuperação e análise de dados espaciais com o intuito de gerar informação geográfica. São 5 os componentes que formam o SIG, desde o operador (uma pessoa), o *hardware* (os componentes da máquina) , o *software* (o programa computacional), os dados e por fim os procedimentos .

A grande vantagem dos SIG sobre outros sistemas de gestão de base de dados, é que ele pode trabalhar através do uso de linguagens de programação orientada por objetos, que nada mais é do que um procedimento baseado não só em funções ou procedimentos, mas também em objetos, os quais poderão ser usados de modo a criar representações abstratas do mundo real, com o intuito de classificar as diversas estruturas espaciais .

Hoje existem diversas opções de plataformas SIG, tanto as pagas, como as gratuitas e de código aberto, os quais são alimentados por uma rede de colaboração internacional de pesquisadores. Para esta pesquisa, toda análise se utilizou do *QGIS* , que se enquadra na segunda categoria de ambiente SIG, e portanto, totalmente gratuito.

O SIG é uma ferramenta vigorosa, capaz de guardar informações espaciais sobre qualquer lugar do espaço. As principais metas de um SIG são: a obtenção e gestão de dados

espaciais, a gestão da base de dados e sua respectiva visualização e o auxílio nas análises espaciais (CONOLLY e LAKE, 2009).

Os SIG trouxeram muitas vantagens nos estudos arqueológicos, tornando as análises muitas mais rápidas. Sanjuán , enumera 6 (seis) vantagens principais na utilização dessas ferramentas, são elas:

• A captura: já que anteriormente era necessária a importação dos dados ambientais de cada sítio e hoje isso é desnecessário, devido ao cruzamento entre as informações arqueológicas e ambientais.

• A integração: nos SIG os dados são facilmente integrados, mesmo sendo das mais diversas fontes

• O georreferenciamento: nestes sistemas os dados são facilmente referenciados dentro do espaço terrestre, o que torna a gestão dos sítios mais fácil.

• A conceituação: Os SIG tornaram os dados arqueológicos mais racionais, de forma que se pode organizá-los como representados por pontos, ou triângulos, ou seja o que se encaixar melhor com o respectivo material a ser apresentado.

• As consultas: tornaram-se muito mais rápidas, devido a não necessidade de se realizar cálculos para responder a perguntas simples, pois os programas já os fazem automaticamente.

• A representação cartográfica: Os SIG possibilitaram uma melhora na qualidade dos mapas arqueológicos, pois fornecem ferramentas capazes de inserir elementos gráficos fundamentais, tais como legenda, escala, norte.

4.2 Sensoriamento Remoto

O sensoriamento remoto é um procedimento técnico que busca obter informações sobre objetos distantes sem realizar contatos físicos com estes, a partir do uso de aeronaves com câmeras acopladas, satélites e/ou uma simples câmera fotográfica ou aparelho de escaneamento a laser. Esses equipamentos buscam medir a energia refletida ou emitida por um determinado alvo, sendo o sol a principal fonte energética utilizada .

Olhando para a Arqueologia, podemos definir, de uma forma mais específica, o sensoriamento remoto como qualquer prática não destrutiva que visa enxergar vestígios enterrados ou não visíveis em determinadas situações. Câmeras aerotransportadas com Lidar, escaneamento a laser e ainda instrumentos geofísicos, são exemplos destas práticas, as quais vem sendo usadas em atividades de campo na Arqueologia nos últimos anos .

O sensoriamento remoto na Arqueologia, é antes de tudo uma técnica auxiliar nos trabalhos de prospecção. Pietroni , afirma que existem três ondas no desenvolvimento da metodologias da prospecção arqueológica. A primeira etapa está ligada aos primeiros estudos, ainda muito presos a noção de antiquarismo, mas que já tinham um olhar sobre a paisagem em que determinados vestígios eram encontrados. A segunda onda, por sua vez, tem um destaque para a utilização de métodos sistemáticos de prospecção e até técnicas de amostragem. Por fim, já no início do novo milênio, com o avanço computacional, técnicas como o sensoriamento remoto e a utilização dos SIG, foram primordiais no avanço das pesquisas em análise espacial.

No cenário da Arqueologia Brasileira, uma das primeiras menções ao uso do Sensoriamento Remoto, vem do trabalho de Chohfi et al , como parte integrante do Simpósio Latino-americano de Sensoriamento Remoto, ocorrido na Argentina. No artigo, os autores discorrem sobre a possibilidade do uso da técnica no auxílio a prospecção arqueológica, e isso numa época onde as principais missões que iriam levar satélites a órbita terrestre, sequer haviam partido.

Dado a essa importância, cabe aqui relatar uma missão de satélite posta em órbita, o SRTM (USGS, 2021), que teve o intuito de gerar imagens sobre a superfície terrestre, e que será usada neste trabalho e servirá de base para a extração de variáveis ambientais.

A missão *Shuttle Radar Topography Mision* (SRTM), foi mandada a órbita terrestre por intermédio da *North America Spatial Agency* (NASA), a bordo da espaçonave *Endeavour* e registrou a superfície terrestre entre os dias 11 e 22 de Fevereiro de 2000. A princípio os dados gerados tinham uma resolução de 30 m, para o território dos EUA, e 90 m, para os demais locais da Terra. Em 2014, porém, a NASA liberou dados com resolução de 30 m, para todo o globo terrestre, abrindo a possibilidade da nova configuração ser usada nesta tese .

4.3 A coleta de dados e a delimitação da área de estudo

Com informações colhidas pela equipe que compôs o Projeto de caracterização e documentação dos sítios rupestres no semiárido de Pernambuco em mãos, foi decidido então, delimitar uma área específica do semiárido pernambucano. Com isso, a região da Bacia Hidrográfica do Pajeú, foi escolhida, por se tratar de um local próximo a áreas já atestadas de passagens de grupos pré-históricos, tais como as cabeceiras dos rios Paraíba, Moxotó e Ipanema.

A delimitação da Bacia Hidrográfica usou como base os dados digitais da Agência Nacional de Águas . Utilizando o *QGIS*, foi feito um "*buffer*", ou seja, uma área com uma distância específica no entorno de uma feição, neste caso a delimitação da própria bacia foi usada como referência. Este contorno apresenta uma espessura de 5km além dos limites do Pajeú, a fim de englobar os sítios que se encontram no entorno imediato da Bacia.

Os dados coletados pela equipe do Projeto de caracterização e documentação dos sítios rupestres no semiárido de Pernambuco, foram inseridos ao *QGIS* e a partir destes, foram obtidas informações tanto de presença, relacionada a ocorrência de sítios arqueológicos com grafismos rupestres, como também de ausências, que tratam de locais visitados pela equipe, porém onde não foram observados nenhum tipo de vestígio.

Estes dados espaciais são inseridos, porém, como ponto no espaço, é necessário, todavia, que estes fossem convertidos em polígonos (representando as áreas dos sítios arqueológicos). Foram então criados polígonos com 100m x 100m, de extensão, os quais englobavam a área abrigada e/ou onde está localizada o bloco com o suporte rochoso e o contexto circundante a este.

Por outro lado, os dados coletados ao longo da prospecção, que indicavam o caminho percorrido pelas equipes de campo, também são inseridos ao SIG, mas estes já foram recebidos como polígonos e passaram a representar as áreas de ausência. Seguindo a metodologia de amostragem proposta por Hobbs no desenvolvimento do *MnModel*, áreas prospectadas muito grandes precisam ser segmentadas. No trabalho por ela desenvolvido, a cada 2km uma nova área prospectada deve ser delimitada, isto, porém é aplicado para todo território do Estado de Minnesota, nos EUA. Como a Bacia do Pajeú é um local cerca de 10

vezes menor, para manter a proporção, será usado aqui um intervalo de 300m para delimitar os quadrados de prospecção.

A base de dados inicial, envolvendo presença e ausência, ainda era muito reduzida e para isso foi necessário criar alguns dados de maneira artificial, utilizando algumas informações prévias. Para realizar tal feito, procurou-se elementos espaciais que pudessem ser incorporados à classe das ausências, como é o caso das áreas urbanas, que são locais já atestados como sem sítios rupestres. Estes foram submetidas a um processo de geração de uma grade de 300x300m, para delimitar dentro das zonas urbanas, novas áreas de ausência, que se somariam as da prospecção.

Os entornos de sítios arqueológicos sem dados de prospecção, também foram usados como base para ampliação da base de dados. É importante explicar que para esse grupo foi realizada a seguinte estratégia. Foi traçada uma área ao redor de um sítio arqueológico sem dados de prospecção, com aproximadamente 2km, e dentro dessa região foram gerados pontos aleatórios, convertidos em seguida para quadrados de 300m de lado, os quais representariam locais sem sítios arqueológicos.

Esse aumento "artificial" da base de dados foi necessário pois dentro dos trabalhos de aprendizagem de máquina, para que se tenham resultados melhores, é preciso ter bases de dados grandes quantitativamente, afinal quanto mais informação, melhor a "máquina" irá aprender.

Neste caso ainda há um outro detalhe, afinal trabalharemos com uma base de dados desbalanceada, já que em qualquer universo da Arqueologia, é muito mais comum lidarmos com locais sem nenhum tipo de vestígio, do que o contrário, e aqui não é diferente, pois teremos muito mais áreas de ausências do que presenças. Neste sentido, baseado em Dumpala, Chakraborty e Kopparapu , é preciso ter uma taxa de desbalanceamento[1] entre as classes majoritárias e minoritárias do conjunto de dados, que nem seja tão alta nem tão baixa, para que os métodos que lidam com este tipo de "problema" surtam mais efeito. Com isto em mente, é

[1] Fração de elementos entre um conjunto minoritário sobre um majoritário (DUMPALA; CHAKRABORTY; KOPPARAPU, 2018)

possível inclusive pensar na estratégia na delimitação das áreas de ausências, para não se criar uma disparidade muito grandes entre os dois grupos. Esta tese trará uma sessão especificamente falando destes métodos mais adiante.

Em face do que foi dito acima, e sabendo que os dados referentes as ausências, previamente coletados, não eram muito abrangentes espacialmente por toda área do Pajeú, essas generalizações foram necessárias para gerar uma base de dados ainda mais diversa, mas que ao mesmo tempo, não caísse no enviesamento provocado pelo pesquisador, já que todo processo foi feito de maneira aleatória. Outras metodologias de aumento dessa base de dados, também foram aplicadas, mas serão explicadas mais à frente neste trabalho, pois já integram a fase prévia de confecção dos modelos.

Por fim, os dados que representam os sítios arqueológicos e que por acaso se sobrepuseram às ausências, foram submetidos a um cálculo de diferença, o qual retira a parte sobreposta pelo outro elemento, impedindo, portanto, que ao calcular as estatísticas zonais, elementos pudessem estar no mesmo local.

Esta primeira parte da preparação dos dados será usada no conjunto que será treinado. Porém, estas informações não correspondem a toda a Bacia do Pajeú, afinal temos áreas em que ainda não se tem conhecimento de sua classificação, como ausência ou presença, a chamada área cinza de dados.

É exatamente a partir dessa área cinza, que serão calculados os novos sítios e ausências ainda não conhecidos, portanto, todo o restante da área de estudo, isso inclui o Pajeú e todo o *buffer* de 5 km, excluindo, logicamente, as áreas já conhecidas. Será então traçada uma grade retangular de 100x100m por toda essa região e assim os dados estarão prontos para serem executados. É importante frisar que esta malha escolhida para o conjunto de execução representará a própria resolução do ráster final.

4.4 Levantando as variáveis da pesquisa

Serão três tipos de variáveis trabalhadas, as de cunho topográfico, socio-ambiental e as de captação de recursos, as quais podem ser consultadas na íntegra nos apêndices desta tese. Dentro do primeiro grupo temos a inclinação, aspecto ou orientação das encostas,

insolação e suas derivações, índice de posição topográfica e o perfil de curvatura. No segundo grupo, temos o índice de visibilidade e a acessibilidade. Por fim, no último grupo, temos as variáveis de distância para os principais cursos d'água e distância para as litologias de acordo com o tipo de rocha.

Essas variáveis foram extraídas tanto a partir dos dados provenientes das imagens de satélites, como também de elementos vetoriais obtidos junto a plataformas governamentais, como é o caso das informações referentes às drenagens, obtidas de uma base de dados da ANA e também da litologia local, obtida junto ao CPRM .

A partir destes dados, foi possível extrair diversas informações utilizando o *QGIS* como interface. Estes atributos obtidos serão submetidos então a cálculos estatísticos e servirão como base para a definição dos produtos de cada MPA gerado. Para melhor compreensão, cada variável do modelo será explicada em detalhes nos tópicos seguintes.

4.4.1 Inclinação

Esta variável, como nome sugere, mostrará o nível de inclinação do terreno em graus. Esta foi obtida utilizando a ferramenta "Declividade", presente no *QGIS* e exige que seja inserido um Modelo Digital de Elevação (MDE). Com isso foram extraídas informações das imagens da missão SRTM .

4.4.2 Aspecto ou Orientação das encostas

Essa variável é obtida usando a ferramenta "aspecto" no QGIS, e demanda apenas os dados de elevação, assim como na inclinação. Ela é medida em graus, e indica, de acordo com os pontos cardeais, para onde estão sendo direcionadas os fluxos. Ao classificar cada um dos intervalos, são diversas possibilidades de direções, partindo do norte geográfico (0°) até o retorno para a mesma posição (360°).

4.4.3 Insolação

A outra variável extraída a partir dos dados de elevação foi a insolação. Utilizando o *software SAGA GIS* foram obtidos os valores da insolação direta, difusa, total, e no período de nascer e pôr do sol. Além disso, a ferramenta permite escolher o período em que os dados serão baseados, portanto, os solstícios de verão e inverno do hemisfério sul, foram usados para isso.

Definindo cada uma das insolações, podemos colocar a "direta", como a insolação medida em um elemento da superfície terrestre perpendicular ao raios solares; a difusa, por sua vez, é a irradiação solar que foi espalhada por atuação da atmosfera; A insolação total é o somatório das duas irradiações citadas acima; Por fim, a insolação nos períodos de nascer e pôr do sol, como o nome sugere, medem a atuação dos raios solares nestas horas do dia .

4.4.4 Índice de Posição Topográfica (IPT)

O IPT é definido pela diferença de um ponto central, presente em um MDE, e a média das células do entorno . Esta métrica pode ser calculada utilizando o algoritmo *Topographical Position Index (TPI)* no *QGIS*, e demanda apenas os dados de elevação.

4.4.5 Perfil de Curvatura

O perfil de curvatura calcula a angulação do terreno de acordo com um intervalo de valores. Com isso o tereno pode ser classificado como convexo, côncavo ou plano . Para calcular esta métrica é necessário apenas o uso da elevação, utilizando para isso o algoritmo "*Slope, Aspect, Curvature*" do *QGIS.*

4.4.6. Distância para os principais cursos d'água

Uma das variáveis que integra o grupo da captação de recursos é a distância para os principais cursos d'água. Foram levantados na base da Agência Nacional de Águas (ANA) , os dados vetoriais que representam as drenagens locais. Alguns processos de modificação destes

dados foram feitos, em que primeiramente foram obtidas as ordens de drenagem seguindo o algoritmo de Strahler, através de uma ferramenta do *QGIS.* Apenas as drenagens de níveis 3, 4 e 5, foram utilizadas.

Após isso, todas essas drenagens principais foram então “rasterizadas”, ou seja, transformadas em ráster. Posteriormente foi criado um mapa de distâncias, que tem o intuito de representar o quão próximo de um elemento ráster estão outros locais dentro de uma margem de limite. Para isso uma imagem que parte do 0 quando está exatamente em cima do que se deseja ser o ponto de origem da medição (as drenagens neste caso), e à medida que se afasta temos o aumento dos valores, até um ponto limite estipulado no cálculo. Para este trabalho utilizamos um limite de 3 km de busca, assim como proposto por Banerjee et al .

4.4.7 Distância para as litologias

O outro representante da categoria da captação dos recursos é a distância para as litologias. O mesmo processo realizado para as drenagens, será feito aqui, já que os dados inicialmente são vetores, portanto será feita uma rasterização e posterior criação do mapa de distâncias. A única diferença para as litologias é que serão criados 3 mapas de distâncias diferentes, um para rochas sedimentares, um para metamórficas e outra para ígneas.

É importante frisar que a criação desse mapa de distâncias terá um intuito de observar duas situações, primeiramente o tipo de suporte rochoso que estão mais próximos a áreas de sítios ou ausências e também, possíveis locais de captação de ocre, que muitas vezes estão relacionados a veios de xistos, filitos, ferricretes, ou seja, ligados a rochas sedimentares ou metassedimentares .

4.4.8 Índice de Visibilidade (IV)

Dentro do grupo das variáveis de cunho social, temos aqui o índice de visibilidade, o qual é calculado a partir de um MDE, e tem o intuito de mostrar o campo visual de uma determinada localização dentro de um dado terreno. Este cálculo demanda um certo poder de processamento,

a depender do tamanho da imagem utilizada e pode ser obtido através do algoritmo "*visibility index",* pertencente ao plugin *QGIS Visibility Analysis* .

4.4.9 Acessibilidade

A última das variáveis, e também integrante das de cunho social, é a acessibilidade, a qual tem o intuito de compreender as movimentações dos povos pretéritos por meio de análise espacial baseada em caminhos de menor custo (CMC) espalhados ao longo do espaço, os quais geram rotas de acesso a um ponto de interesse específico. Para a realização deste procedimento é necessário primeiramente criar uma malha circular de acesso com um 1km de raio nas margens de cada sítio que se tem interesse em compreender acerca das rotas de acesso. Após isso é calculada a chamada superfície de custo acumulada, a qual baseia-se sobretudo na declividade do terreno como superfície de custo e parte desde um ponto de origem até o destino. Por fim, é feito um retrocesso, partindo do destino à origem e são traçados os caminhos propriamente ditos. .

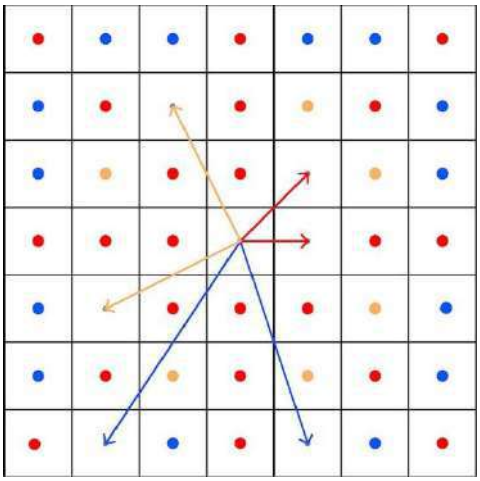
No caso desta pesquisa, porém, utiliza-se uma metodologia baseada em trabalhos como o de Murrieta-Flores e Crabtree et al , que usam os limites da própria área de estudo como as bordas de uma malha de acesso. No primeiro caso, os acessos às áreas de interesse se dão a partir de pontos distribuídos ao largo das margens baseadas sobretudo em características da paisagem favoráveis ao deslocamento de grupos humanos. No segundo caso são utilizados ordem de drenagens de alto nível, para indicar pontos de acesso às áreas de estudo. Com base nisto, utilizou-se aqui esta estratégia de distribuição dos pontos ao longo da Bacia do Pajeú, porém os pontos de acesso foram gerados automaticamente no *Qgis* de acordo com um intervalo fixo de 20 km entre um local e outro.

Para além disso, as análises de acessibilidade não se limitaram ao uso de uma única superfície de custo. Foram criadas redes de caminhos ótimos que ligavam todos os pontos de acesso a todos os pontos de acesso, a chamada abordagem FETE (*From Everywhere to Everywhere*), que segundo White e Barber , é uma grade de pontos de destino/origem sobreposta a uma superfície de custo e cada um dos pontos se interligam. No caso desta pesquisa, como já adiantado, apenas os pontos na borda da área de estudo serão levados em

consideração e tanto a declividade do terreno quanto na insolação direta, serão usados como superfícies de custo.

O cálculo dos CMC desde uma superfície extraída de um ráster passa pelo uso do algoritmo de Dijkstra para caminhos mais curtos . Este passa pela transformação de um arquivo matricial em um grafo com pesos. Entende-se por grafos a ligação entre nós e arestas, neste caso em específico, cada célula ráster será um nó e a ligação entre estes será uma aresta e a mesma terá um peso associado. Cada uma dessas ligações é dada por um movimento parecido com o do xadrez, onde temos o movimento da torre (4 direções), da rainha (8 direções), do cavalo (16 direções) e alguns outros (**Figura 13**).

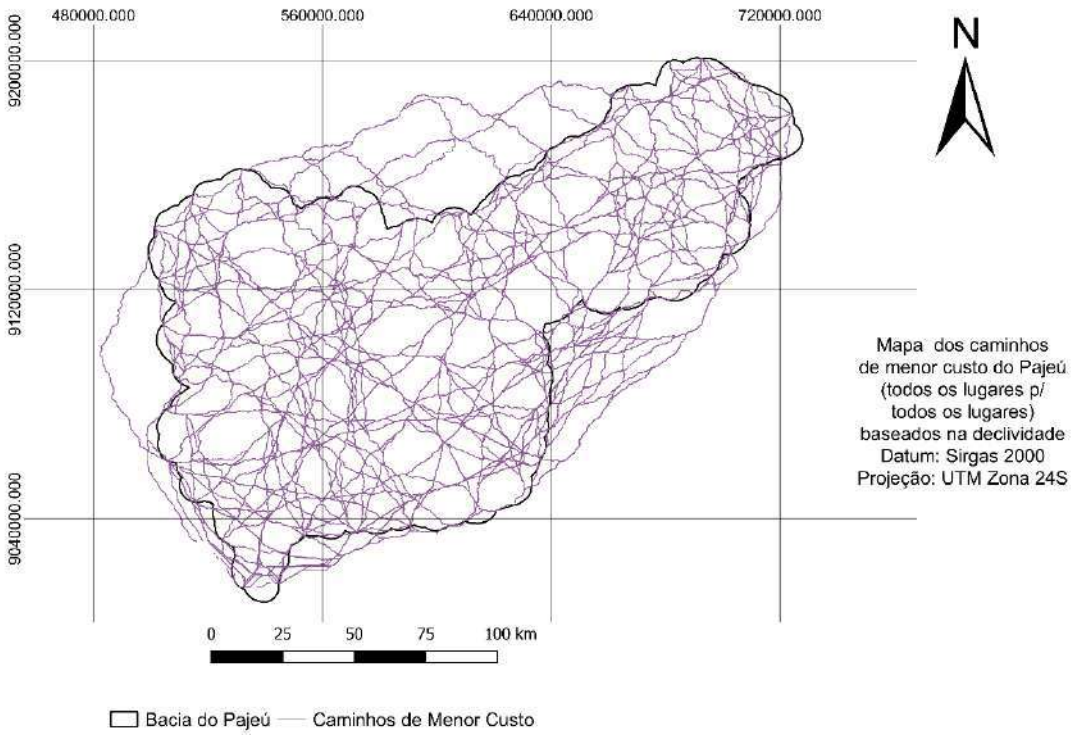
Figura 13: Tipo de movimentos na construção de uma rede de caminhos de menor custo.



Fonte: HERZOG, 2014

No caso desta tese, por questões de limitação computacional, utilizou-se o movimento da rainha, que tem 8 direções diferentes. Essa opção é definida ainda na etapa de criação da superfície de custo acumulado, com o uso da ferramenta *R.walk*, em que será calculada a superfície de custo acumulado através da função de custo de Langmuir e posteriormente, a *R.drain*, para o traçado dos caminhos propriamente ditos (**Figura 14**).

Figura 14: Mapa de caminhos ótimos baseados na declividade como superfície de custo.



Fonte: Elaborado pelo próprio autor (2022)

A partir dos resultados foram gerados quatro produtos derivados, dois mapas de linha de densidade (de 500m de raio por 30m de resolução), que indicam os caminhos ótimos mais usados baseados tanto da declividade como na insolação. Além disso foram criados mais dois mapas de proximidade, que indicam o quão próximo se está dos principais CMC, baseado num raio de busca de 5 km.

4.5 Discutindo as variáveis da pesquisa

Os modelos preditivos, em sua grande maioria, contam com um apelo a uma análise focada no ambiente, o que gera muita crítica quanto ao determinismo ambiental, como já mencionado. Porém, atributos como a visibilidade e acessibilidade, apesar de serem dados obtidos a partir de informações geoambientais, trazem consigo uma conotação social, que atenuam este problema .

Todas as variáveis usadas destes modelos aqui criados, já foram usadas com mais ou menos frequência em outros trabalhos acadêmicos, portanto a utilização destas nesta pesquisa, passa por um embasamento científico atestado. Os modelos que utilizam AM, porém, tem vantagens frente a estatística comum, pois todas essas variáveis serão inseridas nos cálculos, a princípio de maneira igualitária, e seu grau de participação no produto gerado, será acompanhada ao fim da execução dos treinos.

Apesar da literatura trazer esses inúmeros exemplos de atributos para os modelos preditivos, não quer dizer que as variáveis que foram destacadas em uma modelagem, também terão a mesma importância em outra, até porque, cada qual terá seus próprios objetivos e/ou serão usadas para ambientes completamente distintos. Isso é um ponto importante a se falar, um MPA não poderá ser extrapolado para qualquer estudo, é preciso que haja ou uma adequação dos modelos produzidos que atendam uma realidade diferente, ou simplesmente é necessário pensar em novas estratégias que atendam as novas demandas .

Em face disso, os MPA baseados em algoritmos de AM, serão importantes, pois além de informar quais atributos tem maior importância no modelo, também permitem, a partir disto, observar como cada variável se relaciona com sítios arqueológicos estudados e retirar variáveis que pouco contribuem com o modelo, permitindo conhecer ainda mais a relação espacial ser humano/meio.

4.6 Calculando as estatísticas das variáveis

Antes mesmo de inserir as variáveis, de fato, à composição do modelo, é preciso calcular as estatísticas de cada uma delas em relação a cada uma das áreas de sítios e não-sítios. Para isso, utilizando o *QGIS*, foi necessário o uso da ferramenta “estatísticas zonais”, que tem a função de computar os valores de um determinado elemento vetorial em relação a um elemento matricial, ou seja, calcular para cada pixel ou conjuntos deles, quais valores se relacionam estatisticamente com as áreas de ausência e presença previamente delimitadas.

Para cada uma das variáveis citadas no quadro anterior foram calculados cinco métricas estatísticas: a mediana, a moda, o valor máximo, o desvio padrão e a variância. O

motivo de ter escolhido esses diversos parâmetros diferentes para um mesmo elemento é que eles podem fornecer informações diversas.

As métricas de tendência central, medirão o meio do conjunto de dados. Entre as medidas utilizadas acima, a mediana e a moda são integrantes desta categoria. A primeira, porém, indicará o valor central de um conjunto de dados, em que metade dos outros valores são maiores e a outra metade são menores, logicamente. Os valores obtidos através desta implementação são importantes no sentido de identificar *outliers*, valores que fogem do padrão do conjunto. Por outro lado, a moda, indica os valores que aparecem de forma mais recorrente nos dados, sendo importante na verificação de constância nos valores.

Além das medidas usadas acima, também foram calculadas as medidas de dispersão, entre elas o desvio padrão e a variância, cuja importância está no reconhecimento do quão disperso é seu banco de dados. A variância indica o quanto um elemento está distante do ponto central e pode ser calculada através da soma dos quadrados das diferenças de cada observação média dividida pelo número de observações. Por outro lado, o desvio padrão, é expresso pela raiz quadrada da variância .

Por fim, foi aplicado o cálculo do valor estatístico máximo, com o intuito de observar os valores mais elevados dentro do conjunto de dados. Como estas métricas fazem gerar mais 5 atributos de uma única variável, temos ao final dos cálculos 120 variáveis a serem trabalhadas como mostrado no quadro presente nos anexos desta tese.

4.7 A construção do MPA baseado em algoritmos de AM

Ao final de toda a fase de preparação dos dados, e relacionamento das variáveis com cada área a ser trabalhada, é chegada a hora de iniciar a aplicação dos MPA propriamente ditos. Nesta pesquisa será usado o *Python*, uma linguagem de programação orientada por objetos, que é definida como clara e poderosa, segundo o próprio site do projeto .

O *Python* é hoje uma das linguagens de programação mais usadas no mundo, sobretudo no cenário da IA . Na ciência, como um todo, ela vem conquistando seu espaço a cada ano, mas nem sempre foi assim. Foi apenas no início dos anos 2000, com o fortalecimento

da comunidade ScyPy, que o uso desta entrou de vez dentro dos estudos científicos . O projeto está hoje na sua 3ª versão (*Python3*) e muito foi mudado e incorporado, para facilitar o uso e tornar a linguagem mais parecida com o próprio inglês escrito.

A linguagem é formada por uma série de elementos, dos quais os módulos, são um dos principais. Estes são definições e códigos escritos em *Python* e salvos em arquivos do tipo *.py*, os quais servirão como comandos específicos para realizar determinada ação. O conjunto de diversos módulos é chamado de biblioteca e a linguagem *Python* tem diversas delas, uma para cada função, as vezes até mais de uma para uma mesma função, com algumas diferenças .

4.7.1 Exportando os dados para o Python

O primeiro passo para a criação de um MPA baseado em algoritmos de AM, é importar estes dados base até uma interface capaz de executar a linguagem de programação desejada. Os dados exportados do *Q*GIS vêm com o formato *.csv*, para tanto é necessário o uso de uma biblioteca *Python* específica, neste trabalho se usou o *pandas*. Esta tem entre outras funções ler e gerar arquivos tabulados em *csv, xlsx, txt,* entre outros .

Ainda nessa primeira fase dentro do *Python*, um outro ponto importante a se fazer foi a normalização de toda a base de dados, que nada mais é do que o ato de deixar todas as colunas numéricas referentes a cada uma das 125 variáveis, em uma mesma escala de valores e para isso foi utilizada uma outra biblioteca, a *Numpy* .

4.7.2 Estruturando a AM

Antes mesmo de adentrar no próximo passo desta metodologia, cabe aqui informar algumas nomenclaturas e categorizações importantes. O escopo que envolve a AM é bem vasto, porém podemos dividi-la em quatro tipos de aprendizado: o supervisionado, semi-supervisionado, o não-supervisionado e o por reforço .

Neste trabalho, o enfoque maior será na utilização do aprendizado supervisionado, em que os dados alimentados pelo algoritmo contêm as informações ou rótulos, de toda a base previamente levantada, e usa dessas informações para aprender e obter respostas para os dados não rotulados

. Explicando em outras palavras, quando temos dados de sítios e de não-sítios conhecidos, usamos desta informação obtida para saber a respeito de locais que ainda não foram visitados e, portanto, não tem classificação alguma.

Os algoritmos de classificação e regressão são os grandes representantes dessa categoria de aprendizado, sendo o primeiro composto de rótulos de classes finitas, associados geralmente a algum tipo de característica nominal podendo apresentar duas classes ou múltiplas. Um grande exemplo da utilização desses algoritmos no dia-a-dia é na distinção entre e-mails classificados "spams" ou não, ou também no resultado de exames, mostrando se um paciente está acometido de alguma doença .

Por sua vez, o segundo, trata da previsão de valores reais e contínuos, sobretudo números, sendo um exemplo comum do cotidiano a previsão no preço de um imóvel, levando em conta as características físicas deste, como tamanho, cor e número de aposentos .

4.7.3 Divisão dos conjuntos de dados

Após esta breve introdução a respeito do tipo de aprendizagem que será aqui usado, os olhos se voltam aos passos metodológicos da pesquisa. Após a exportação e normalização do conjunto de dados, é chegado o momento de dividir a base de informações em 3 parcelas: os conjuntos de treino, teste e validação.

Para a criação desses conjuntos citados é preciso, porém verificar as "sementes" (*seeds*), que nada mais são do que valores numéricos que são usados para certificar que tanto o treino quanto a validação, terão composições idênticas todas as vezes que forem executados. Em um modelo ideal, o uso de valores diferentes para essas sementes não deve gerar grandes alterações nos resultados avaliativos, se isto ocorre há claramente um problema com os dados trabalhados.

Nesta tese, utilizamos 30 valores de sementes diferentes a fim de atestar a confiabilidade dos conjuntos de dados gerados, avaliando estatisticamente os resultados através de valores de precisão, *recall* e acurácia para cada um dos algoritmos utilizados.

Esta separação dos conjuntos evita ou atenua os chamados sobreajuste (*overfitting*) ou sub-ajuste (*underfitting*). O primeiro faz com que os dados treinados dentro de uma base sem separação dos conjuntos se comportem de uma maneira positiva, já que o modelo só lidará com dados conhecidos. Quando novas informações externas à base são inseridas, esta passa a

cometer grandes erros, já que apenas decorou as informações dadas a ele. O segundo, por sua vez nem chega a ser testado, já que o treino indica uma falta de relações entre as variáveis escolhidas. Tudo isso provoca um enviesamento no resultado final .

Um outro risco a se levar em conta quando se trabalha com modelos de AM, e que já foi levantado quando explanado a respeito da necessidade de se trabalhar com bases de dados grandes, é com a quantidade de dados usados na parcela de treino e a generalização destes. É preciso que as informações sejam suficientes para realização de uma modelagem confiável, e que os dados sejam o mais abrangentes possível, caso contrário o modelo não se comportará de maneira adequada, causando um ruído amostral (*sampling noise*) quando a amostra é muito pequena, ou um enviesamento amostral (*sampling bias*) quando há uma generalização muito ampla dos dados (GÉRON, 2019).

A divisão dos conjuntos nesta tese seguirá os seguintes valores: para a parcela de treino, que é a maior dentre todas, foram usados 70% dos dados. Esse grupo de informações tem a função de fornecer os dados necessários para que o modelo se adeque à novos exemplos, que porventura venham a ser a acrescidas ao modelo. Em resumo, como o nome sugere, é a partir dele que serão extraídos os padrões que farão a "máquina aprender" (ABU-MOSTAFA; MAGDON-ESMAIL; LIN, 2012; BURKOV, 2019; GÉRON, 2019).

Os outros dois conjuntos que compõem a base de dados do modelo, e tem um tamanho muito mais reduzido, quando comparados aos dados de treino, são a validação (15%) e o teste (15%). O primeiro tem a função de ser usado para facilitar na escolha do algoritmo de aprendizagem a ser usado e buscar um melhor desempenho do modelo. Já o segundo, por sua vez, tem a função de pôr o modelo em prática, com dados totalmente alheios ao conjunto de treino. (ABU-MOSTAFA; MAGDON-ESMAIL; LIN, 2012; BURKOV, 2019; GÉRON, 2019).

4.7.4 Conjunto de dados desbalanceados

Após feita a separação dos conjuntos, é realizada uma outra etapa, que visa lidar com um problema enfrentando em conjuntos de dados que tem uma classe muito maior que a outra, em termos quantitativos. Como tratamos aqui de uma dicotomia ausência/presença,

teremos, pois, muito mais locais sem nenhum tipo de vestígio arqueológico, do que o contrário, o que caracteriza o *database* desta pesquisa como desbalanceado .

Para lidar com tal "problema" é preciso realizar um tipo de implementação estatística conhecida como reamostragem, que pode ser aplicada com o uso da biblioteca *Imbalanced_Learn* , do *Python*, diretamente sob o conjunto de treino ou com o uso de algoritmos já preparados para lidar com tais situações. Aqui neste trabalho usaremos dois tipos diferentes de reamostragens, uma chamada de *oversampling* ou sobreamostragem, caracterizado por um balanceamento a partir da classe minoritária, buscando igualá-la a majoritária e o outro, por sua vez, chamado de *undersampling* ou subamostragem, cujo objetivo é atuar na classe majoritária com o intuito de equipará-la a minoritária .

Dentre as reamostragens usadas podemos destacar primeiramente o *KMeans + Smote*, em que primeiro é aplicado um método de agrupamento chamado *kmeans*, que terá a função de encontrar algumas correlações com as diferentes classes existentes (ausência/presença), tentando melhor agrupá-las. Em seguida, é executada a *Synthetic Minority Oversampling Technique* ou Técnica de Sobreamostragem Sintética Minoritária (SMOTE) , que busca igualar as classes minoritárias com as classes majoritárias, por meio da interpolação entre várias instâncias semelhantes que estão próximas .

O segundo método utilizado, o *SMOTE*, simplesmente, é igual ao anterior, com uma única diferença, já que a aplicação do *Kmeans*, não é utilizada. Por fim, diferentemente dos dois primeiros, o terceiro método de reamostragem aqui utilizado, a Abordagem de Amostragem Sintética Adaptativa ou *Adaptatative Synthetic Sampling Approach* - ADASYN , insere amostras sintéticas de classes à menor representatividade, as quais são mais difíceis de serem aprendidas pelos métodos de Aprendizagem de Máquina, já que como falado, amostras muito pequenas podem causar o enviesamento amostral.

4.8 Os algoritmos

Este trabalho utilizará o método de aprendizagem supervisionada, focado na utilização de algoritmos de classificação. Dentre os principais tipos de métodos classificatórios podemos destacar a Classificação Linear, a Regressão Logística, K-Vizinhos próximos (*K-Nearest*

Neighbours-KNN), Máquina de Vetores Suporte (*Support Vector Machine-SVM*), Árvores de Decisão (*Decision Trees*) e Redes Neuronais Artificiais (RNA) .

Neste trabalho serão usadas a regressão logística, uma RNA e dois modelos baseados em árvores, a árvore de decisão e a floresta aleatória, que é um conjunto de diversas árvores de decisão. Além destes, serão aplicados dois algoritmos (floresta aleatória balanceada e *bagging* balanceado), também baseados em árvores, mas que já são preparados para lidar com dados desbalanceados e, portanto, não necessitam que seu conjunto de treino seja reamostrado. Por fim, o último modelo será uma mistura de todos estes algoritmos, um modelo *ensemble*.

4.8.1 Regressão linear

Apesar de não ser aqui utilizado, a regressão linear é uma das bases para todo o aprendizado estatístico. Como relatado no capítulo 2 desta tese, ela já vinha sendo desenvolvida desde o século XIX e é a base para o desenvolvimento de outros métodos de aprendizagem de máquina.

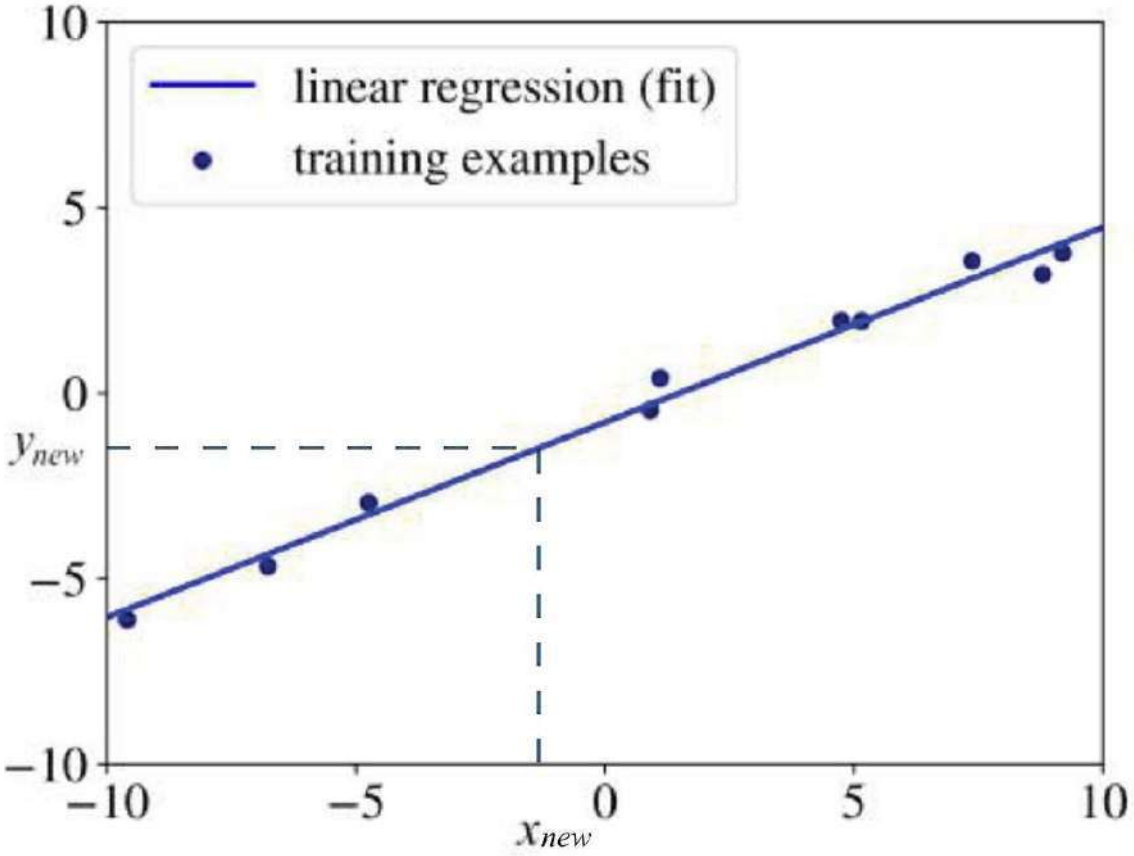
A regressão linear, como nome indica, é um algoritmo de regressão e portanto visa encontrar valores específicos para um determinado objetivo. Para compreender essa implementação estatística e como ela vai auxiliar no entendimento da configuração de uma regressão logística, precisamos primeiramente entender o que é uma função linear, a qual é a base para todo o modelo do tipo linear. Esta função pode ser definida pela equação abaixo:

$$\hat{y} = w_0 + w_1 x$$

Onde $\hat{y}$ é a hipótese, ou valor predito, w_0 e w_1 são os parâmetros ou pesos, em que o primeiro representa o ângulo de inclinação da reta e o segundo é o intercepto, que cortará a reta vinda do eixo x. Por fim , o *x* representa a variável do modelo .

Para ilustrar, a figura abaixo (**Figura 15)** mostra como funciona na prática um modelo de regressão linear e como ele procura se ajustar ao máximo através dos dados de treino que são fornecidos. É importante pontuar, que como mostra o gráfico, dados que não tenham comportamento linear não conseguem ser preditos com modelos deste tipo, pois levariam a enviesamentos e produtos errôneos.

Figura 15: Modelo de regressão linear



Fonte: BURKOV, 2019

Modelos como os de regressão linear, que contém os chamados algoritmos baseados em otimização[1], demandam o cálculo de uma função de erro ou de custo, com o objetivo de obter o menor custo do modelo. Uma das formas mais simples de obter esse valor é através do cálculo do erro quadrático médio (*mean squared error* ou MSE), mostrado na equação abaixo:

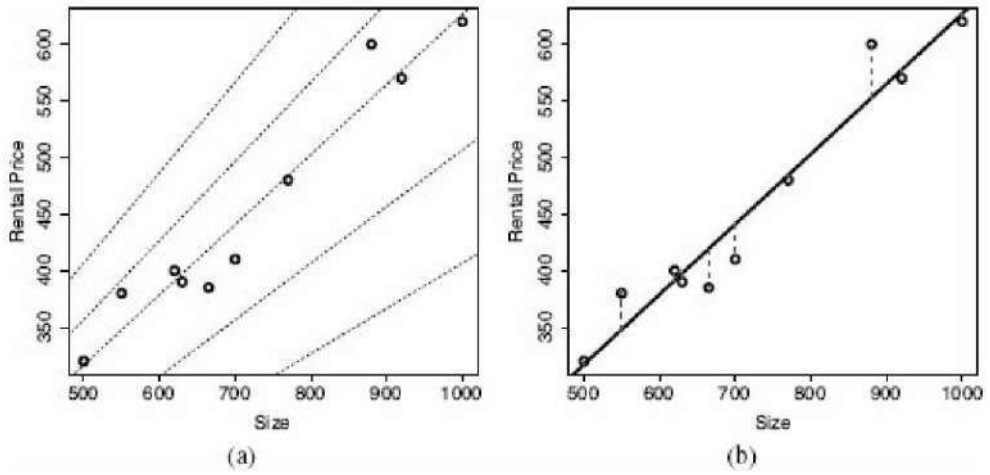
$$C(w_0, w_1) = \frac{1}{m}\sum_{i=1}^{m} \hat{y} - y^{(i)2}$$

Nessa função, o *C* equivale ao custo ou valor de MSE, *m* é a média da função, *i* é a iésima parte de um valor atrelado tanto aos preditores quanto ao alvo, $\hat{y}$ é a hipótese, já descrita

[1] Algoritmos como regressão linear, regressão logística e redes neuronais artificiais, são entendidos como algoritmos baseados em otimização, pois buscarão o erro mínimo na função.

anteriormente e *y* é o valor observado na reta da função. A representação do erro e das diversas retas de otimização podem ser vistas na figura abaixo (**Figura 16**) .

Figura 16: gráficos de linhas de ajuste do modelo de regressão linear (à esquerda) e percepção dos erros associados a linha melhor ajustada (à direita)



Fonte:

4.8.2 Regressão logística

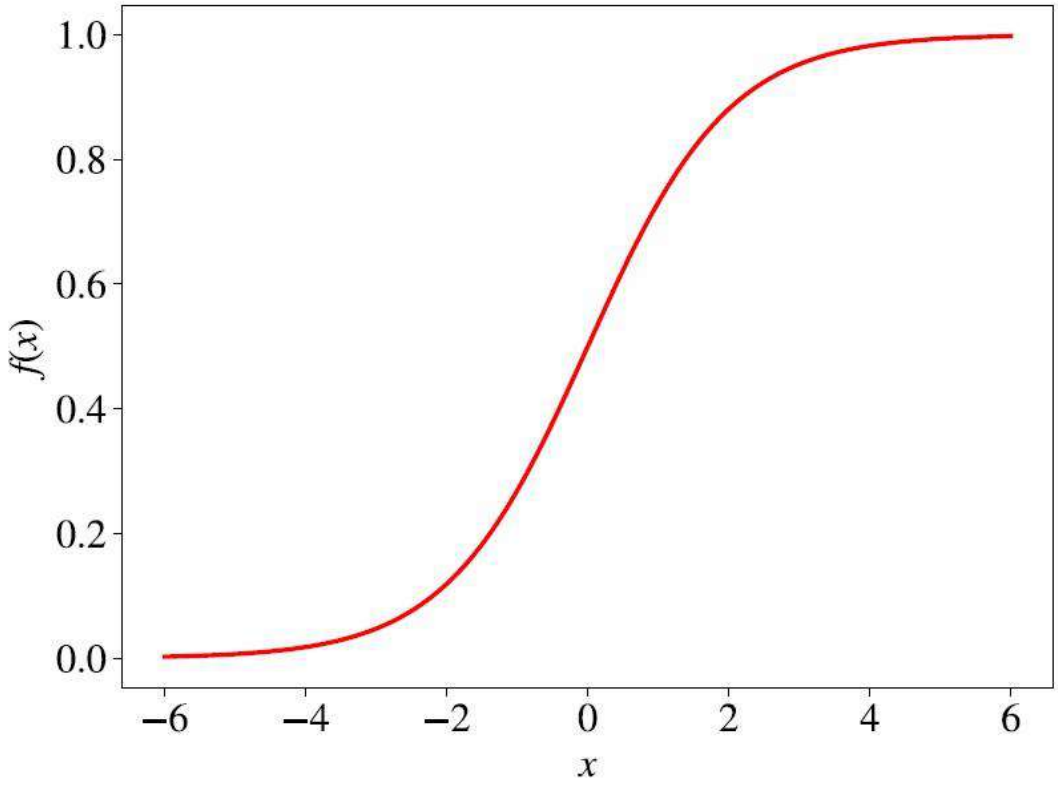
A regressão logística é um dos algoritmos de classificação usados nesta pesquisa. Antes de mais nada, é preciso ficar claro que apesar do nome, carregar o termo “regressão” este pode ser utilizado como classificador logístico, ao passo que trabalha dentro de uma perspectiva linear, assim como a regressão linear. Seu resultado, porém, é expresso através de uma probabilidade, com valores que variam entre 0 e 1 . Quanto mais próximo de 0, teremos os valores negativos, neste caso as ausências. Do contrário, quanto mais próximo de 1, indicam uma maior probabilidade de termos locais com sítios de grafismos rupestres. Essa é uma configuração simples, levando em consideração apenas duas classes, e claramente pode ser ampliada para outros contextos, que não vem ao caso deste trabalho.

O modelo de regressão logística, para ser expresso como proposto acima, com valores que variam entre 0 e 1, necessitam de expressões particulares que diferem da regressão linear. Uma das formas de se expressar essa configuração é através da função sigmoide ou

função logística, que pode ser expressa pela seguinte equação e tem esta representação gráfica (**Figura 17**):

$$f(x) = \frac{1}{1 + e^{-x}}$$

Figura 17: Função logística ou sigmoide



Fonte: BURKOV, 2019

Interpretando os dados da equação acima, temos que a função logística é igual razão de 1 sobre 1 adicionado ao número de *Euler* ou logaritmo natural, representado pela expressão e^{-x} (BURKOV, 2019).

Esta é a representação da função logística, para representar o modelo de regressão logística propriamente dito, os elementos da função linear são então inseridos a equação, sendo expresso desta maneira:

$$\hat{y} = \frac{e^{w_0 TX_0 + w_1 TX_1 + \cdots + w_n TX_n}}{1 + e^{w_0 TX_0 + w_1 TX_1 + \cdots + w_n TX_n}}$$

A equação acima mostra como a função logística relaciona a estrutura matemática do modelo de regressão linear. Esmiuçando os termos, se tem novamente a razão entre o número de *Euler* elevado a estrutura da função linear, dividido por 1 adicionado ao número de Euler elevado a função linear.

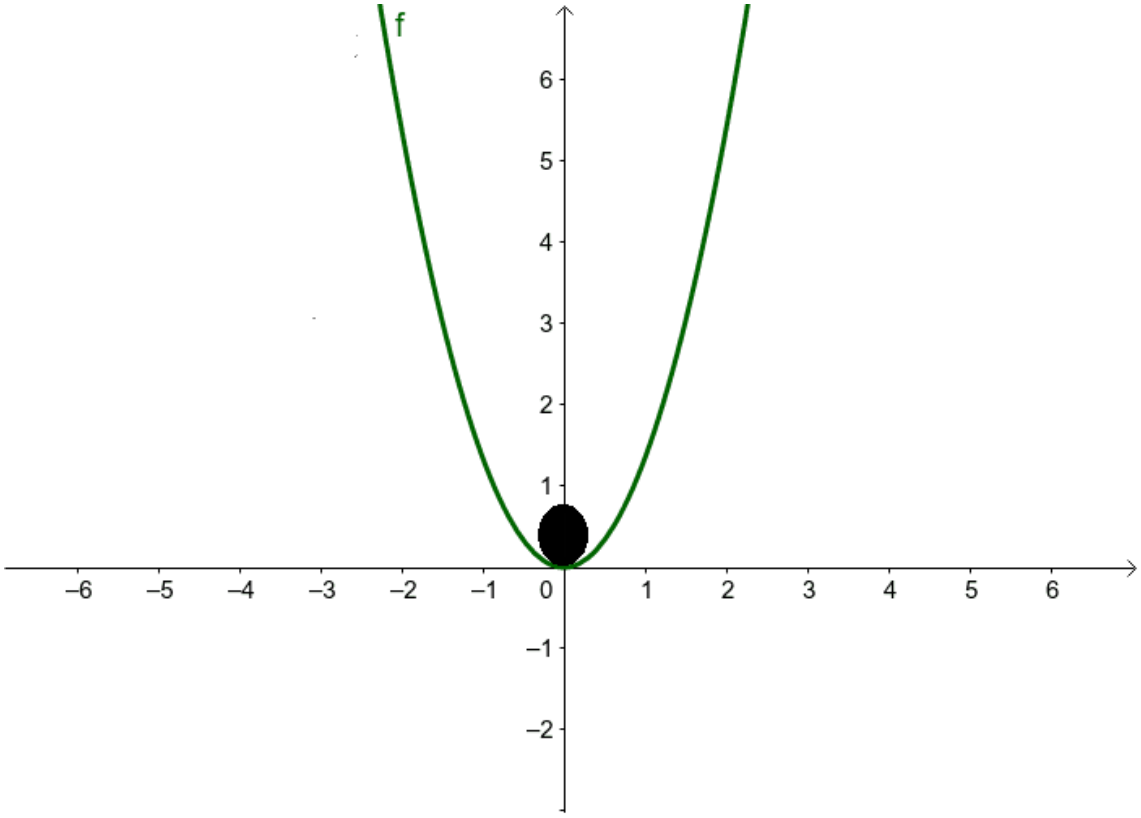
De forma detalhada, temos que w_0 é o valor inicial da matriz de peso e X_0 é uma variável criada, de modo a poder igualar a quantidade de valores da matriz de pesos. Essa implementação é feita, seguindo as regras da álgebra linear, já que multiplicação entre os dois fatores *w* e *X*, só acontece quando a matriz que carrega os pesos é transposta, permitindo a multiplicação entre duas matrizes do tipo 1x0 e 0x1, ou seja, que tem número de linhas de uma, igual ao número de colunas da outra.

Como no nosso exemplo de trabalho não lidaremos apenas com uma única variável, os outros valores de *w* e *X*, representam cada qual, um peso e uma variável diferente, até se esgotarem todas as variáveis do modelo, representado na equação pela expressão $w_n X_n$.

Para saber os valores dos pesos que acompanharam cada uma das variáveis, é preciso utilizar um outro tipo de recurso, os chamados processos de otimização. Um dos mais usados na AM é o gradiente descendente. O modelo de regressão logística aqui utilizado, usa uma variação deste, chamada de gradiente descente estocástico . Para ser ainda mais específico, podem ser encontradas algumas variantes deste tipo de algoritmo, aqui optou-se por utilizar a SAGA , integrante da biblioteca *Scikitlearn*.

O gradiente descendente é uma função que buscará o local de menor valor do erro. A variante estocástica, apenas é diferente no sentido de o cálculo ser feito apenas em uma parcela do conjunto, o que torna o processo mais rápido e menos custoso computacionalmente, já o processo matemático, é basicamente o mesmo. Se pensarmos em uma função quadrática, do tipo $(ax + b)^2$, esta terá uma configuração gráfica como mostrada na figura abaixo (**Figura 18**), e, portanto, a intenção do gradiente descendente é descobrir quais parâmetros levam ao mínimo valor dessa curva.

Figura 18: Representação de uma função quadrática com o ponto de erro mínimo



Fonte: Elaborado pelo próprio autor (2021)

Para obter os valores do gradiente descendente é preciso descobrir os valores para cada um dos pesos que acompanham as variáveis usadas no modelo, de acordo com a formulação abaixo:

$$w_0 \coloneqq w_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \hat{y} - y^{(i)} x_0$$

$$w_1 \coloneqq w_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \hat{y} - y^{(i)} x_1$$

$$w_2 \coloneqq w_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} \hat{y} - y^{(i)} x_2$$

$$\vdots$$

$$w_n \coloneqq w_n - \alpha \frac{1}{m} \sum_{i=1}^{m} \hat{y} - y^{(i)} x_n$$

A função acima é fruto de um processo de dedução matemática que envolve algumas etapas, mas o importante a se compreender nela é que os valores de w_n :=representam o gradiente descendente de cada peso; w_n representa o valor dos parâmetros de uma variável na posição "n", ou seja, o processo tem que ser feito para todas as variáveis do modelo; α representa a taxa de aprendizagem do modelo, que indicará quantos "passos" por vez, serão dados para se chegar ao ponto ótimo, como indicado na figura 11; a parcela central dessa equação ($\sum_{i=1}^{m} \hat{y}$- $y^{(i)}$ é apenas o cálculo do custo, anteriormente demonstrado; por fim, o erro será multiplicado pelo valor de cada variável, representado aqui por x_n.

Com todos esses processos, é chegado o momento de usar a implementação estatística que é capaz de obter os critérios de otimização do modelo de regressão logística. Ao contrário da regressão linear, que busca minimizar os riscos usando o MSE, aqui será usada a chamada probabilidade máxima ou *maximum likelihood*, que pode ser expressa da seguinte forma:

$$E_{\hat{y}}(w) = \frac{1}{N}\sum_{i=1}^{N} ln\left(1 + e^{-y_n w^T x_n}\right)$$

Olhando a equação acima fica evidente que a dedução matemática para se chegar aos valores acima vem da estrutura da função sigmoide e da estrutura da função linear. O *N* representa todos os exemplos que serão treinados no modelo; *ln* é o logaritmo neperiano ou natural, usado aqui para inverter a função exponencial advinda da sigmoidal. O restante da equação nada mais do que a própria estrutura da função logística já apresentada anteriormente neste tópico.

4.8.3 Árvores de decisão e florestas aleatórias

Dentre todos os métodos de classificação, o mais simples e que apresentam uma melhor compreensão e é de fácil interpretação, são os modelos baseados em árvores, ou mais comumente chamados de árvores de decisão. Segundo Bishop (2011), estes são selecionados a partir de uma sequência de decisões, que será iniciada de um nó inicial ou raiz (atributo central) e a partir destes, surgirão, na maioria das vezes, duas ramificações (galhos), que configuram o

processo binário (quando há apenas duas classes). A cada novo nódulo gerado, dois outros galhos surgirão, até as possibilidades se esgotarem.

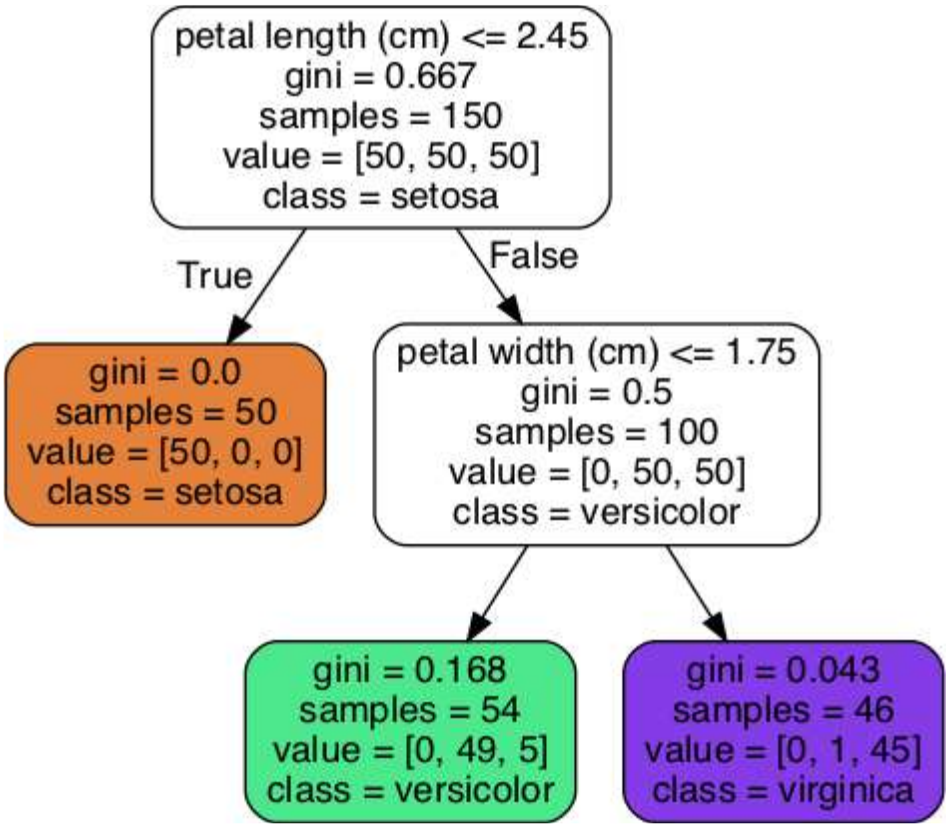
Existem para tanto, várias estruturas baseadas em árvores, as mais utilizadas são a CART (Classification and Regression Trees), ID3 e C4.5 . O *scikitlearn* usa uma versão otimizada do CART em seus modelos baseados em árvores, a qual será aqui definida.

A estruturação de uma árvore de decisão que utiliza o algoritmo CART segue uma série de passos. Primeiramente, terá que ser definido qual atributo será utilizado na raiz da árvore, para tanto é preciso realizar testes estatísticos (ex.:Coeficiente de correlação de Pearson) que verão qual atributo se encaixa melhor no nó raiz. A partir daí são gerados os "descendentes" deste nó, que serão ramificações das árvores. Em cada novo nódulo criado são feitos testes a fim de selecionar o melhor atributo para cada posição da árvore .

Quando se toma como exemplo uma das bases de dados mais usadas como teste, o conjunto de dados Iris, que traz informações sobre uma espécie de flor e para tanto, caso seja encontrada, é possível classificá-la, como tal. Para isso, inicia no nó raiz ou (profundidade 0) e a partir daí são feitas "perguntas", como por exemplo, o comprimento da pétala é menor que 2 cm? Caso positivo, uma nova ramificação é gerada e o nodo folha, é criado. Este nodo só é denominado desta forma, porque não haverá mais descendentes dele, já que ele atendeu a demanda da pergunta.

Por outro lado, caso outra flor seja avistada, e sua resposta seja negativa, ou seja, a pétala é maior que 2cm, um novo nodo é gerado, porém este não será mais um nodo folha, afinal não se atendeu a demanda inicial. Com isso novas características ou variáveis serão inseridas a árvore, até que as possibilidades se esgotem ou até que a árvore atinja a profundidade máxima, que pode ser indicada no *Scikit-learn* A estrutura dessa árvore pode ser vista na figura abaixo (**Figura 19**).

Figura 19: Árvore de decisão Conjunto de Dados Isis



Fonte: GÉRON, 2019

O algoritmo CART, que o *scikit-learn* utiliza, faz uso de uma implementação estatística para selecionar qual variável e limiar serão usados em cada nodo, e para isso é feita uma divisão do conjunto de treino em duas partes, usando uma única variável e um único limiar e escolhe o par (limiar/variável) mais puro, segundo a equação de custo abaixo:

$$J(X, t_x) = \frac{m_{esquerda}}{m} G_{esquerda} + \frac{m_{direita}}{m} G_{direita}$$

Nesta equação temos em *G* a medida da impureza do lado direito e esquerdo da árvore. Já *m* é o número de instâncias dos lados direito e esquerdo da árvore. O valor de *G*, porém só pode ser obtido, mediante um outro cálculo, que dentro do *scikit-learn*, por padrão, é feito o *Gini impurity*, porém é possível selecionar a entropia, um outro tipo de implementação, que será aqui escolhida. Ela pode ser representada pela seguinte equação :

$$G(t) = -\sum_{i=1}^{l} \left(P(t=i) x log_S \left(P(t=i) \right) \right)$$

Nessa equação *P(t=i)* representa a probabilidade do resultado da escolha aleatória de um elemento t ser do tipo i, l é número de diferentes tipos no conjunto de dados, e S é uma base logarítmica qualquer. O sinal de "menos" no início da notação é apenas para converter os valores negativos que surgirão a partir do cálculo do logaritmo .

Os modelos de árvores de decisão são os mais interpretáveis, têm custos computacionais baixos, porém há algumas limitações que os acompanham. A principal delas é fato destes serem muito sensíveis a mudanças no conjunto de treino, que podem resultar em problemas de sobreajuste, ou seja, o acréscimo de novos dados ao modelo faz com que estes não tenham bons desempenhos.

Os problemas como os sobre e sub ajuste, já relatados anteriormente, podem ser amenizados com a implementação dos chamados modelos comitê. Segundo Bishop (2011) estes funcionam considerando a média das predições de um conjunto de modelos individuais, em que cada um contém erros atrelados diferentes, uns contém variâncias maiores ou menores e/ou enviesamento maior ou menor, a junção de todos eles vai diminuir o erro associado.

Uma das formas de diminuir o sobreajuste e o enviesamento atrelado aos modelos de árvores de decisão, é criar as chamadas "florestas aleatórias" (*random forest*), estas, funcionam da seguinte maneira: a partir do conjunto de dados de treino do modelo, serão criados diversos outros conjuntos de forma aleatória, que servirão de base para a construção de diversas árvores de decisão. A cada etapa do sequenciamento da árvore, são inseridas variáveis de forma aleatória, evitando assim uma forte correlação entre estas. Ao final do processamento de todas as árvores, será feita uma votação pela maioria (voto majoritário), chegando assim ao alvo do modelo .

O modelo de floresta aleatória utiliza o paradigma de *bagging*, na maioria de seus casos. Este funciona da seguinte maneira: são criadas várias cópias diferentes do conjunto de treino e feitas árvores de decisão usando cada amostra do treino. Essa amostragem é realiza através de um método chamado amostragem por substituição ou *sampling with replacement*, o qual consiste em iniciar a partir de um conjunto vazio e ir selecionando cópias aleatoriamente

dos exemplos do conjunto de treino. Esse processo encerra quando o valor de amostras for igual ao valor de cópias .

Num classificador de Floresta aleatória, apenas uma etapa é adicionada ao paradigma de *bagging*, pois a cada nova divisão do conjunto é feita uma inspeção das variáveis. Esse processo evita que se criem árvores com forte correlação, já que variáveis com alto poder de predição, tendem a ser replicadas em mais exemplos do que outras e a *random forest* acaba diminuindo esse problema .

4.8.4 As redes neuronais artificiais (RNA)

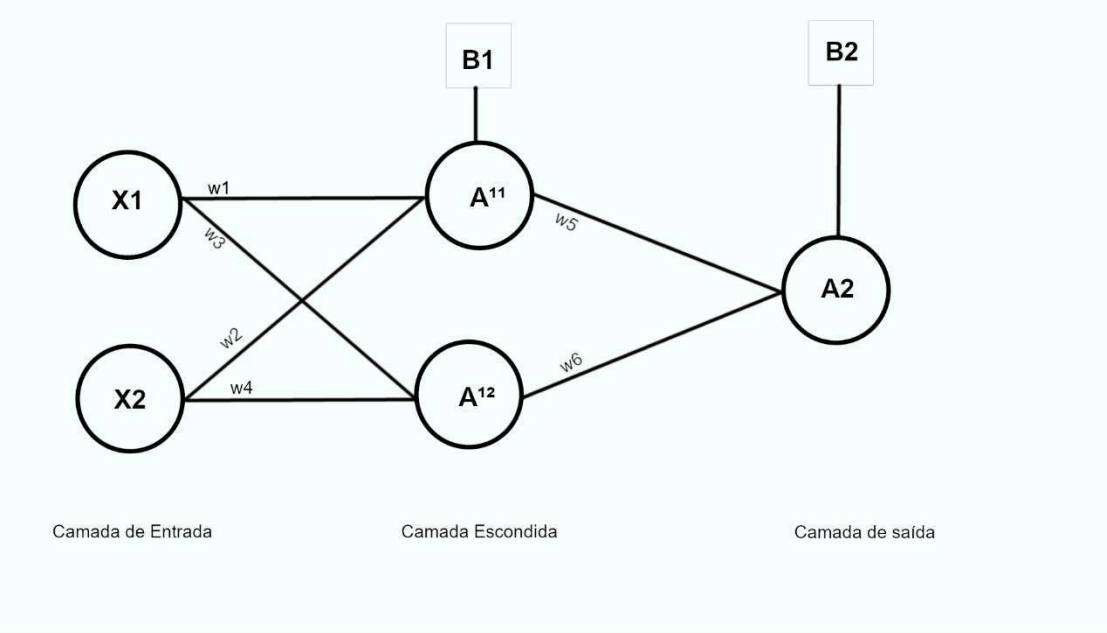
Os modelos de redes neuronais artificiais, foram um dos primeiros a serem desenvolvidos no ramo da AM, como já relatado anteriormente nesta tese. Neste espaço do trabalho cabe defini-las e explicá-las de uma forma mais específica. Ao contrário dos outros modelos de classificação aqui trazidos, a RNA terá uma arquitetura diferente, pois além das camadas de *input* que contém as informações de cada variável e seus pesos e a camada de saída, em que se localiza a hipótese, esta terá camadas internas entre essas duas, conhecidas como camadas ocultas.

Matematicamente, as RNA têm uma estrutura muito similar a regressão linear, ou a modelos de regressão logística, porém esta trabalhará com funções dentro de outras funções, ou seja, é uma composição de funções, como é mostrado na equação abaixo:

$$y = f_{RNA}(x) = f_3\left(f_2\left(f_1(x)\right)\right) \rightarrow f_i = g_i(w_i z + b_i)$$

Observando a notação acima, fica claro que a base matemática de uma função do tipo ax + b, continua sendo implementada, apenas sendo antecedida por uma função de ativação *gi*. O i representa o índice da camada, e identificará cada uma das camadas calculadas na porção interior do modelo; Os valores de wi e bi seguem representando os parâmetros ou pesos do modelo e serão definidos pelo gradiente descendente.

Figura 20: Rede Neuronal Perceptron Multicamadas com uma camada escondida.



Fonte: Elabora pelo próprio autor (2022)

Para ficar mais claro, a imagem acima (**Figura 20**) ilustra como está estruturada um modelo de RNA em multicamadas chamadas *perceptron*. Analisando a imagem fica claro que os dados x são as variáveis de entrada, seguidos por uma camada, que nesse caso contém dois *perceprtrons* ou neurônios e cada um dos inputs trazidos de início, se ligam a todos os neurônios. Passando para a camada seguinte, temos mais uma estrutura formada por mais um neurônio, que se comunicarão com todos os neurônios anteriores e por fim, gerarão um resultado ou hipótese.

Tendo sua RNA estruturada é preciso saber, como em todas as funções de otimização, o erro associado a ela. Assim como na regressão linear, os pesos são obtidos de maneira aleatória em uma primeira propagação através da rede. Com isso, ao final do processo, ao se chegar na camada de saída, o erro associado ainda é muito grande, tornando-se necessária a aplicação de um algoritmo específico que buscará o menor erro possível numa função de ativação tal como a sigmoide.

Dentre os possíveis algoritmos, o mais largamente conhecido é o chamado de *back propagation*, o qual é dividido em duas fases, uma é a propagação e a outra é a retro propagação.

Em outras palavras a propagação segue a rede no sentido *input/output* e retro propagação segue o sentido oposto. Esse processo é feito até que se chegue no valor mínimo do erro.

4.8.5 Outros algoritmos

Para além dos algoritmos citados acima, que irão precisar serem submetidos a um método de reamostragem, posterior a aplicação deles, serão utilizados também, outros dois algoritmos que já são preparados para lidar com um conjunto de dados desbalanceado. Estes serão a floresta aleatória balanceada (FAB) e o *bagging* balanceado, ambos podem ser encontrados como módulos da biblioteca *Imbalanced Learn.*

A FAB , que ao contrário da versão tradicional de florestas aleatórias, não cria conjuntos aleatoriamente, e utiliza uma outra metodologia, extraindo um conjunto de exemplos da classe minoritária, e a mesma quantidade da classe majoritária, a fim de equilibrar o conjunto de dados. Esta parcela dos dados é usada na etapa de treino, realizando um método de subamostragem.

No caso do chamado *bagging* balanceado, o qual também é baseado em árvores, assim como o *bagging* tradicional e no caso modelo aplicado neste trabalho, também é usada uma técnica de subamostragem, ou seja, ele diminui os dados majoritários afim de igualá-los aos minoritários .

4.9 Avaliando o modelo

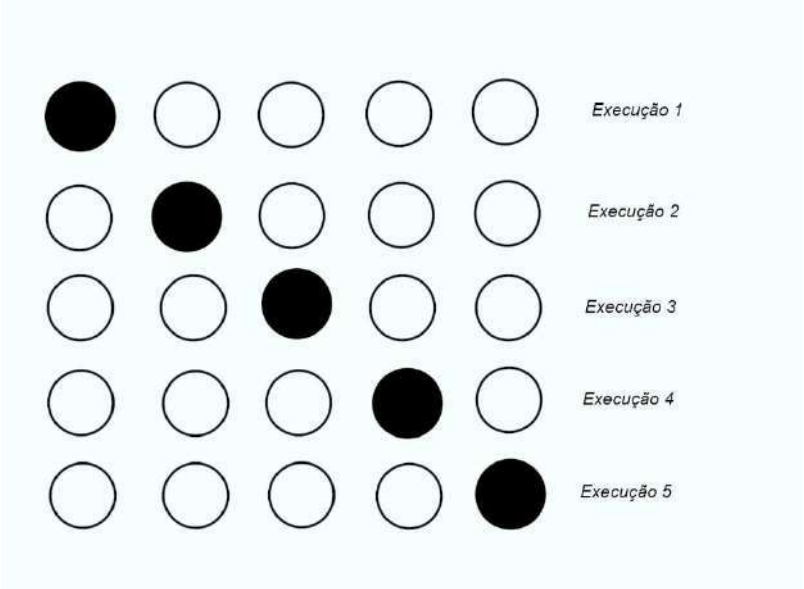
Uma outra importante etapa dentro da construção de um modelo de AM são os métodos de avaliação. Estes podem ser divididos em duas frentes, primeiramente os métodos de projeção de experimentos avaliativos, terão a função de calcular a melhor estimativa em relação ao desempenho de um modelo, auxiliando na estratégia de construção do mesmo, verificando para isso os dados utilizados, sobretudo nos conjuntos de teste e validação do modelo. A outra parcela dos métodos avaliativos terá como função primordial, avaliar a eficiência dos modelos propriamente ditos, após terem sido colocados em prática .

Para que os dados de treino e teste utilizados no modelo sejam confiáveis, é preciso realizar algum tipo análise que comprove o bom desempenho dos mesmos. Uma forma possível de se chegar a estes pode ser através dos métodos de projeção de experimentos avaliativos, com destaque para validação cruzada. Esta terá funcionalidade em modelos que contam com dados limitados, sobretudo no conjunto de validação, já que a escassez de informação pode levar a ruídos nas estimativas de predição .

Dentro dos algoritmos de regressão logística, RNA, árvore de decisão e floresta aleatória que serão aplicados utilizando módulos do *Scikit-learn*, existe uma opção de buscar os melhores parâmetros a serem usados para a execução do modelo com o conjunto de validação. Esta ferramenta é chamada de *GridSearchCV*, em que será feita uma procura pelos melhores parâmetros usando para isso a validação cruzada como método.

Esse método permitirá que uma proporção *S-1/S* de todo o conjunto de dados disponível seja usado para treino e todo o dado disponível seja utilizado para se medir o desempenho do modelo. Quando os dados são mais limitados é recomendável considerar S = N, em que N é o número total de pontos de dados, e a técnica de *leave-one-out*, é aplicada. Está funcionará da seguinte maneira (**Figura 21**), em N diferentes vezes, o modelo será treinado usando todos os pontos de dados, excetuando um, que será usado para previsão, que vai pegar o conjunto de dados por completo e a partir daí dividir aleatoriamente as parcelas de treino e teste. .

Figura 21 - Esquema representando o processo de *leave-one-out.*



Fonte: Elaborado pelo próprio autor (2022)

Este processo será feito até todos os dados serem calculados, levando em consideração o conjunto de dados como um todo. Porém, há uma outra forma de se utilizar da validação cruzada, usando para isso apenas a parcela de dados referente a validação e fazer o mesmo processo anterior, mas em um conjunto de informação mais reduzido, que será o caso neste trabalho.

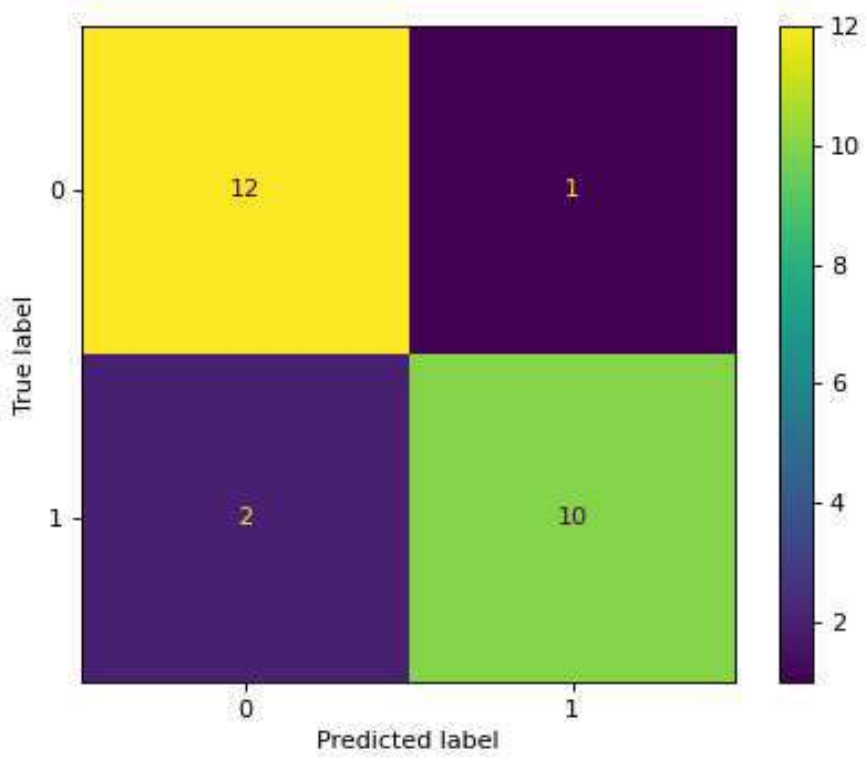
Para além disso, os algoritmos de classificação passam por um tipo de avaliação do quão bem-sucedido é o modelo. Existem diversos testes que podem medir tal parâmetro, tais como: a Matriz de confusão; acurácia, acurácia de sensitividade-custo, precisão, *recall*, área abaixo da curva ROC (AUC/ROC), entre outros (BURKOV, 2019).

Estas avaliações seguirão a busca por quatro valores de dados principais, os verdadeiro-positivos (VP), os verdadeiro-negativos (VN), os falso-positivos (FP) e os falso-negativos (FN). Em ordem, o primeiro representa os valores que foram previstos como verdadeiros e, de fato são. O segundo contém os dados previstos como falsos e que são realmente falsos. O terceiro mostra os valores previstos como verdadeiro, mas que, na realidade, são falsos. E por fim, os últimos representam os valores previstos como falsos, mas que são verdadeiros .

Em aplicações médicas, por exemplo, a separação desses valores e a obtenção de resultados com uma maioria de verdadeiro-positivos e verdadeiro-negativos é muito importante, já que na prevenção ou na confirmação de que um paciente sofre de algum tipo de problema cardíaco, ou algum outro tipo de doença grave, pode ser rapidamente identificada ou descartada e a partir daí as estratégias por parte do profissional de saúde serão traçadas .

Diante disso pode-se expor cada um dos métodos supracitados, tendo na matriz de confusão o exemplo mais simples, que nada mais é do que a organização dos resultados obtidos após execução do modelo, em uma matriz de NxN, onde serão apresentados os valores de VP, VN, FP e FN e a partir daí o pesquisador poderá traçar estratégias para melhorar seu modelo. Como exemplo, o *Scikit-learn*, gera matrizes como as mostradas na figura abaixo (**Figura 22**).

Figura 22: Exemplo de matriz de confusão



Fonte: SCIKITLEARN.ORG, 2022

Lendo este exemplo para o caso deste trabalho, teríamos no canto superior esquerdo, os valores previstos corretamente como áreas de ausência e no canto superior direito os valores previstos incorretamente como sítios arqueológicos, quando na verdade eram áreas de ausência. Na parte inferior da matriz, temos do lado esquerdo os valores previstos incorretamente como ausências e do lado direito os previstos corretamente como sítios.

A forma mais usada para se acessar a eficiência dos modelos com dados desbalanceados é através da precisão, do *recall*, e do Fβ ($\beta = 2$(medida de F2), os quais podem ser obtidas pelo uso das equações abaixo:

$$Precisão = \frac{VP}{VP+FP} \qquad Recall = \frac{VP}{VP+FN} \; F2 = \frac{(1+\beta^2)XPrecisãoXRecall}{\beta^2 XPrecisão+Recall}$$

.

Como a maior preocupação neste trabalho é saber onde estão os sítios arqueológicos, este voltará os olhos com mais afinco para o *recall*. Como afirma Brandsen , a Arqueologia busca tarefas de alto recall, já que a preocupação maior é obter o máximo de informações relevantes possível, mesmo que no meio haja algo irrelevante. No caso desta pesquisa, os erros na classe negativa (ausência), medidos pela precisão, tem baixo impacto no modelo, pois os riscos de se classificar um local que não tem vestígio, de forma errada, ou seja, classificá-lo como uma área de presença, não causará danos ao patrimônio arqueológico, pelo contrário, este fator amplia a margem de erro de uma possível execução do modelo no mundo real.

4.9.1 Avaliando as variáveis do modelo

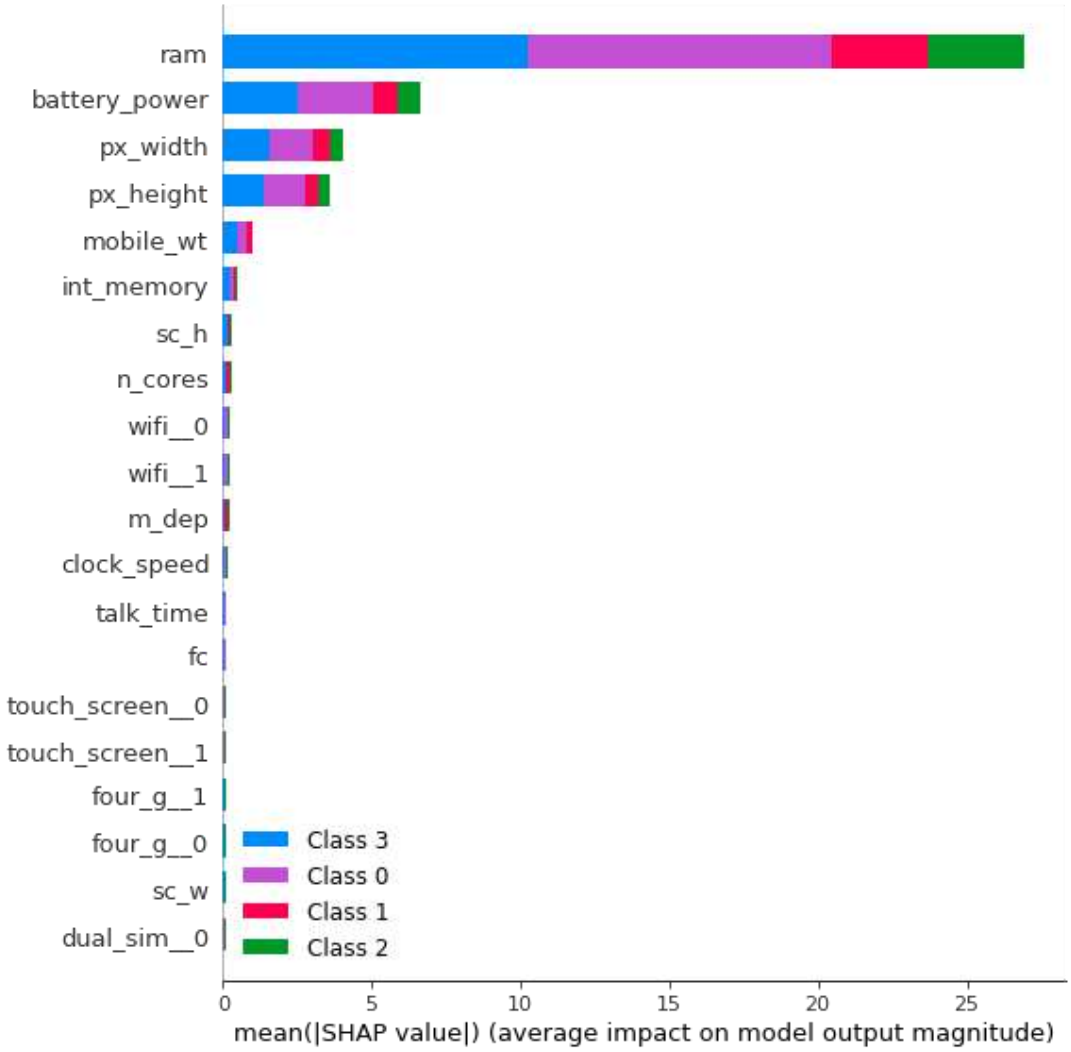
Um dos grandes problemas dos modelos de AM é a dificuldade encontrada para interpretá-los da melhor forma, sobretudo quando falamos de Florestas Aleatórias e RNA, são os chamados *black boxes* . Existem algumas opções no mundo computacional para sanar esses problemas, mas muitas delas são limitadas e fazem que quem deseja melhor compreender o modelo gerado, não saiba qual a melhor escolha a ser utilizada para determinado tipo de algoritmo. Pensando nisso Lundberg e Lee desenvolveram o *SHAP* (*SHapley Additive exPlanations*), como forma de unificar diversos métodos para explicação de modelos de AM.

Para criar os mais diversos modelos explicativos, o *SHAP* se vale das ideias base da Teoria dos Jogos. Primeiramente é preciso pois, transformar as variáveis em jogadores e as funções do modelo de AM nas regras do jogo. Assim como nos jogos, onde um jogador pode ou não se juntar a uma partida, os atributos de um modelo, podem ou não contribuir ou estar inserido em um determinado modelo. Dito isto, toda vez que tivermos os valores das variáveis, é sabido que as mesmas participaram do cálculo, do contrário, temos a certeza que elas não participaram e portanto não interferiram no modelo .

Para se chegar aos valores *SHAP*, os quais indicarão as importâncias das variáveis de cada modelo, o caráter aditivo destes é levado em consideração. Com isso, o *input* de todas as variáveis do modelo será somado pela diferença entre os valores esperados e os valores obtidos ao se executar o modelo de AM .

Existem diversos algoritmos específicos para cada tipo de modelo de AM, no caso da regressão logística é possível gerar histogramas com as porcentagens de cada uma das variáveis usadas no modelo, usando para isso o algoritmo explicativo *linearExplainer*. Com este é possível obter os valores *SHAP* de qualquer modelo linear. Para além disso, o algoritmo trabalha com amostras do conjunto de dados, o que o torna computacionalmente simples. Os valores *SHAP* podem ser expressos de diversas formas, aqui nós utilizaremos algo como mostrado no gráfico abaixo (**Figura 23**), através do uso de gráficos de barra horizontais.

Figura 23: Representação gráfica dos valores *SHAP*



Fonte: GUPTA, 2020

Para os modelos do tipo RNA perceptron multicamadas, estes podem ser obtidos seguindo um outro algoritmo chamado de *KernelExplainer*. Este, ao contrário do anterior pode trabalhar com amostras menores do *dataset*, porém ainda assim exige mais poder de processamento, o que demanda mais tempo para executá-lo. Aqui utilizamos uma amostra de 100 valores do conjunto de treino.

No caso dos modelos de árvores de decisão, floresta aleatória, e quaisquer outro baseado em árvores, é usado um outro algoritmo específico, o *TreeExplainer*. Este também

exige um poder computacional baixo, assim como *linear*. Os modelos de Floresta aleatória e FAB foram interpretados a partir desse algoritmo.

4.9.2 Interpretando as variáveis

Após ter acesso a importância das variáveis, é possível tornar as informações ainda mais interpretáveis com o uso do *QGIS*. Depois todos os dados terem sido exportados ao SIG, é possível relacionar as principais variáveis do modelo, às áreas de presença e assim saber quais elementos melhor contribuem para o achado de sítios arqueológicos com grafismos rupestres.

Com base nas informações adquiridas nos gráficos *SHAP*, é possível criar produtos ráster ainda mais específicos, como por exemplo, em um modelo hipotético, em que o aspecto seria uma das variáveis em destaque. É possível então reclassificar as informações referentes a orientação da encosta, baseado nos pontos cardeais e com isso gerar um novo mapa com valores que estarão dentro de intervalos de ângulos, já que neste caso teremos valores entre 0º e 360º.

Com isso em mente, seria possível calcular as estatísticas do modelo para as áreas positivas, baseando-se no novo mapa reclassificado, o qual nos daria informações muito mais precisas a respeito de quais elementos estão melhor contribuindo para se achar sítios arqueológicos com grafismos rupestres.

4.10 Testando o modelo em campo

Após toda a fase de execução e avaliação do modelo, que é feita toda através de implementações estatísticas, é preciso levar todo o conhecimento adquirido, para o campo. Como já foi discutido nesta tese, um dos grandes problema dos MPÀ é que muitas vezes eles sequer são testados em campo, baseando apenas nas métricas matemáticas, ou ainda, quando são posto em práticas, locais com média e baixa probabilidade são preteridos .

Nesta tese utilizamos uma amostragem diferente. Primeiramente, ainda no ano de 2020, foram construídos os primeiros modelos, com os dados possíveis e apenas utilizando como base para extração de variáveis, além das variáveis geológicas e de proximidade com

cursos d’água, foram usados os dados do projeto SRTM, para extração das outras variáveis ambientais e socioambientais.

Esse primeiro modelo foi muito importante, já que serviu de base para construção de estratégias para a ida a campo durante o mês de junho de 2020. Nessa campanha foram visitados 4 áreas diferentes entre os municípios de Carnaubeira da Penha, Mirandiba e Floresta, onde foram visitados locais com baixa, média e alta favorabilidade e os dados colhidos em campo, mesmo que não contivessem sítios arqueológicos, foram inteiramente registrados, em formato de traçado e posteriormente exportado para o *QGIS*.

Todos esses dados serviram para a ampliação da base de dados, a qual foi preparada novamente, passando por todas as etapas relatadas no início desta metodologia, a fim de criar novos modelos. Essa situação representa bem o ciclo de construção de um bom MPA. Não adianta apenas construir toda uma estrutura computacional e não aplicar diretamente o seu produto ao mundo real, pois é só com vários destes ciclos que as informações vão sendo colhidas e as modelagens vão ficando cada vez melhores.

É importante frisar que o uso de ferramentas como a AM e o emprego de todas estas etapas, não é a solução para todos os problemas relacionados aos MPA, que por se tratar de uma modelagem, são apenas formas de melhorar o poder explicativos dos eventos históricos, sem que isso seja uma verdade absoluta .

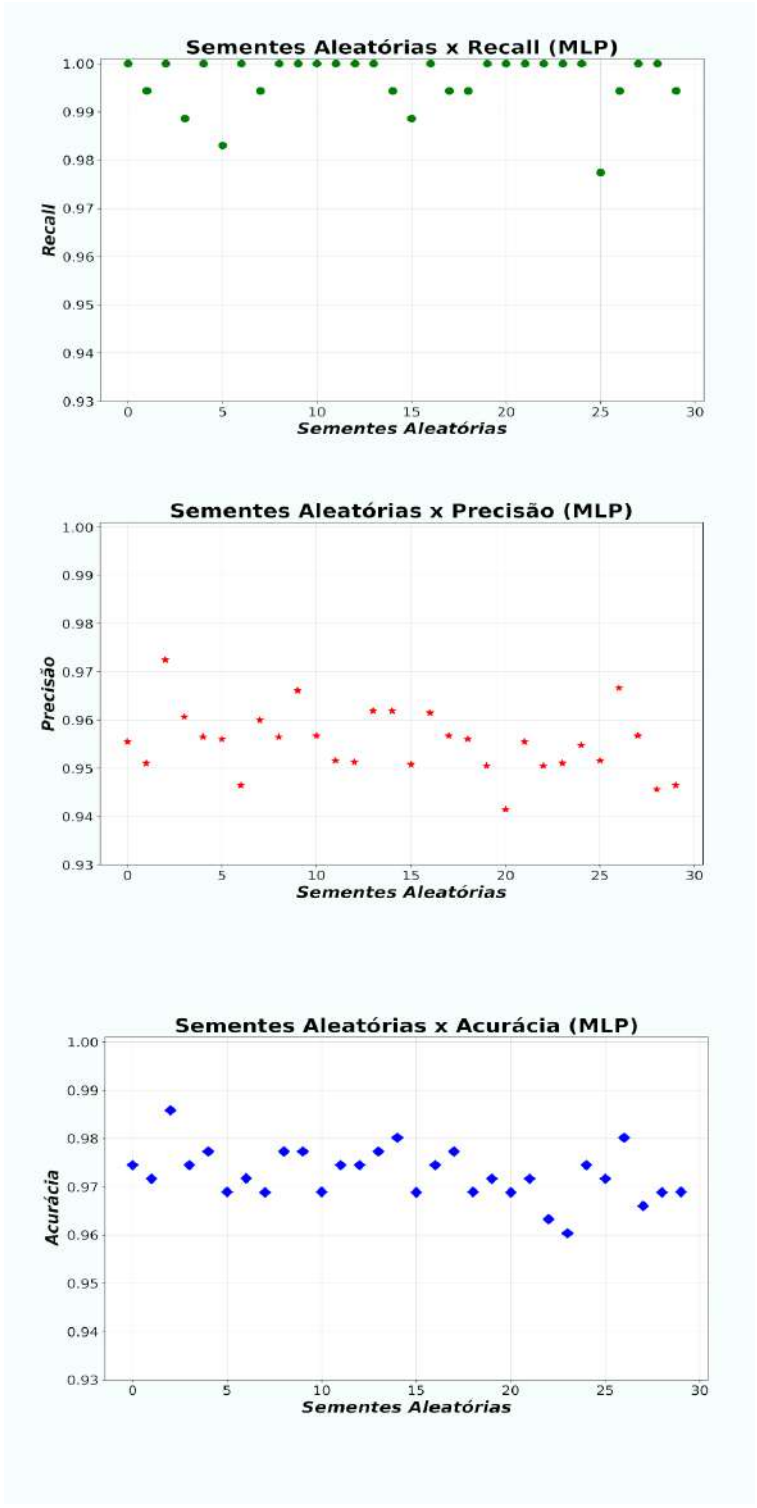
5. MPA_PAJEÚ_0.1: O PRIMEIRO MODELO

Este capítulo trará os primeiros resultados com os MPA produzidos para os sítios de registro rupestre da Bacia do Pajeú. Os produtos gerados a partir destes modelos puderam servir de guia para uma primeira ida ao campo durante o mês de junho de 2020, de modo a testar a aplicabilidade do modelo e também coletar mais informações *in loco*.

Em primeiro lugar, é importante mostrar como se deu a avaliação das sementes aleatórias utilizadas na separação dos conjuntos de dados do modelo. Os resultados podem ser observados no gráfico abaixo, que abarcam apenas os resultados referentes ao modelo de Rede Neuronal Artificial Perceptron. O restante deles podem ser conferidos nos anexos desta tese.

Os valores apresentados nos gráficos baixo (**Figura 24)** nos mostram que a variação na escolha das sementes em nada interfere no resultado final dos modelos, já que os valores de *recall*, precisão e acurácia permanecem quase inalterados quando comparados, o que nos permite utilizar qualquer semente sem nenhum tipo de problema. Essa análise foi feita para todos os algoritmos e o restante dos gráficos encontra-se nos apêndices desta tese.

Figura 24: Gráficos com a distribuição dos resultados referente as avaliações de cada uma das sementes aleatórias para o modelo de RNA.



Fonte: Elaborado pelo próprio autor (2022)

5.1 Os primeiros resultados

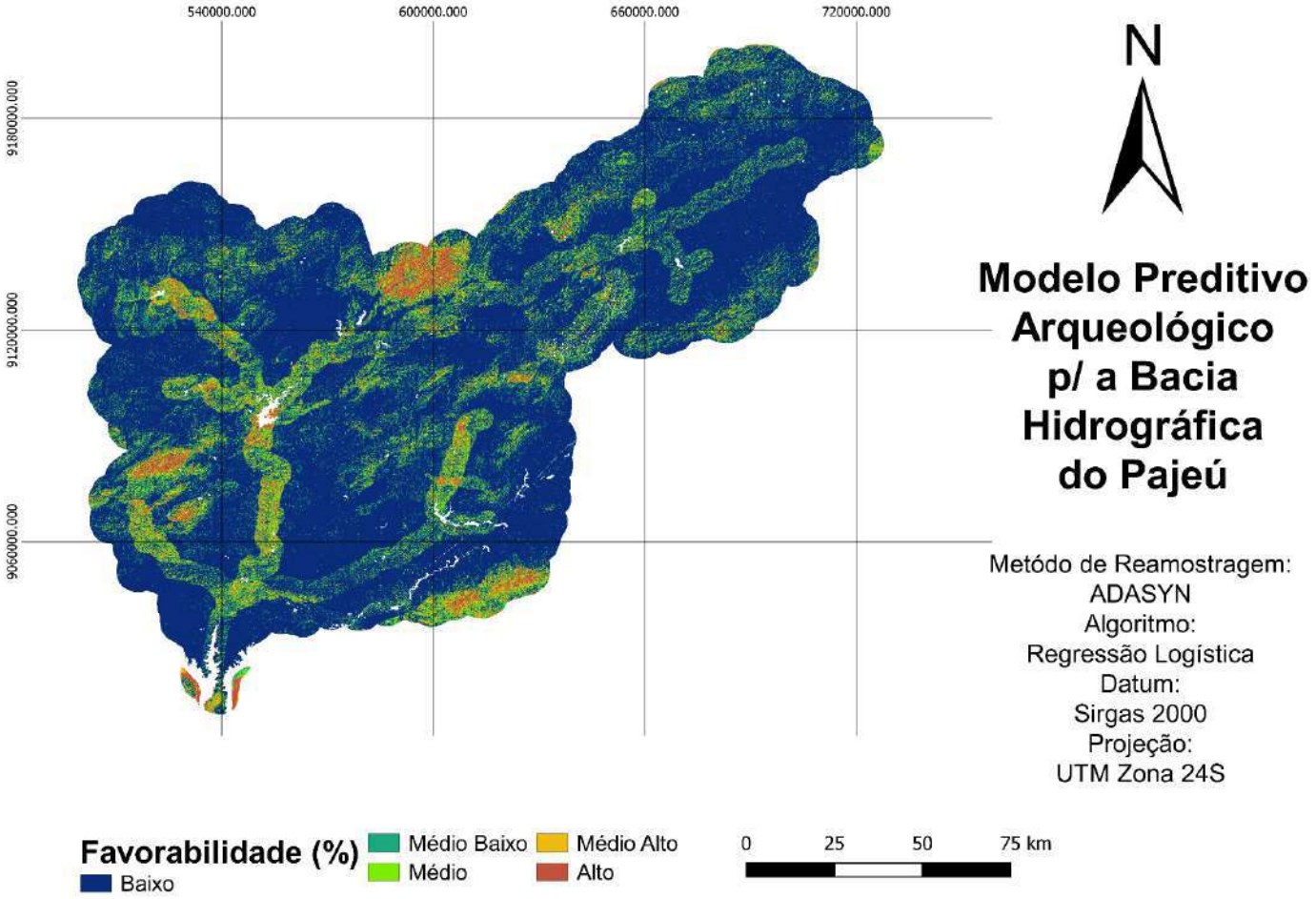
A área escolhida para o primeiro teste do modelo, abrangeu sobretudo os territórios dos municípios de Mirandiba, Carnaubeira da Penha e Floresta. É importante relatar que o todo o trabalho de campo foi feito durante a pandemia de Covid-19 e por isso delimitou-se um perímetro mais reduzido, o que tornou possível analisar locais com baixa, média e alta favorabilidade de se encontrar sítios arqueológicos.

O local também foi selecionado pois nele estão localizados dois territórios indígenas pertencentes a dois povos diferentes, os Atikum, na Serra do Umã e o os Pankará, na Serra do Arapuá. Devido a essa realidade territorial, somado ao fato da presença de alguns sítios rupestres da região e achados de vestígios cerâmicos e líticos, escolheu-se a área para serem aplicados os primeiros testes práticos do modelo.

Os modelos aplicados nesta primeira fase, seguiram todos os passos relatados na metodologia, com exceção de dois pontos. O primeiro ponto é que aqui não foi executado alguns algoritmos, tais como a RNA e a Árvore de Decisão, que só viriam a ser aplicadas num segundo modelo. Além disso, a variável “acessibilidade” não havia sido ainda incorporada ao escopo do trabalho e por isso também não foi inserida.

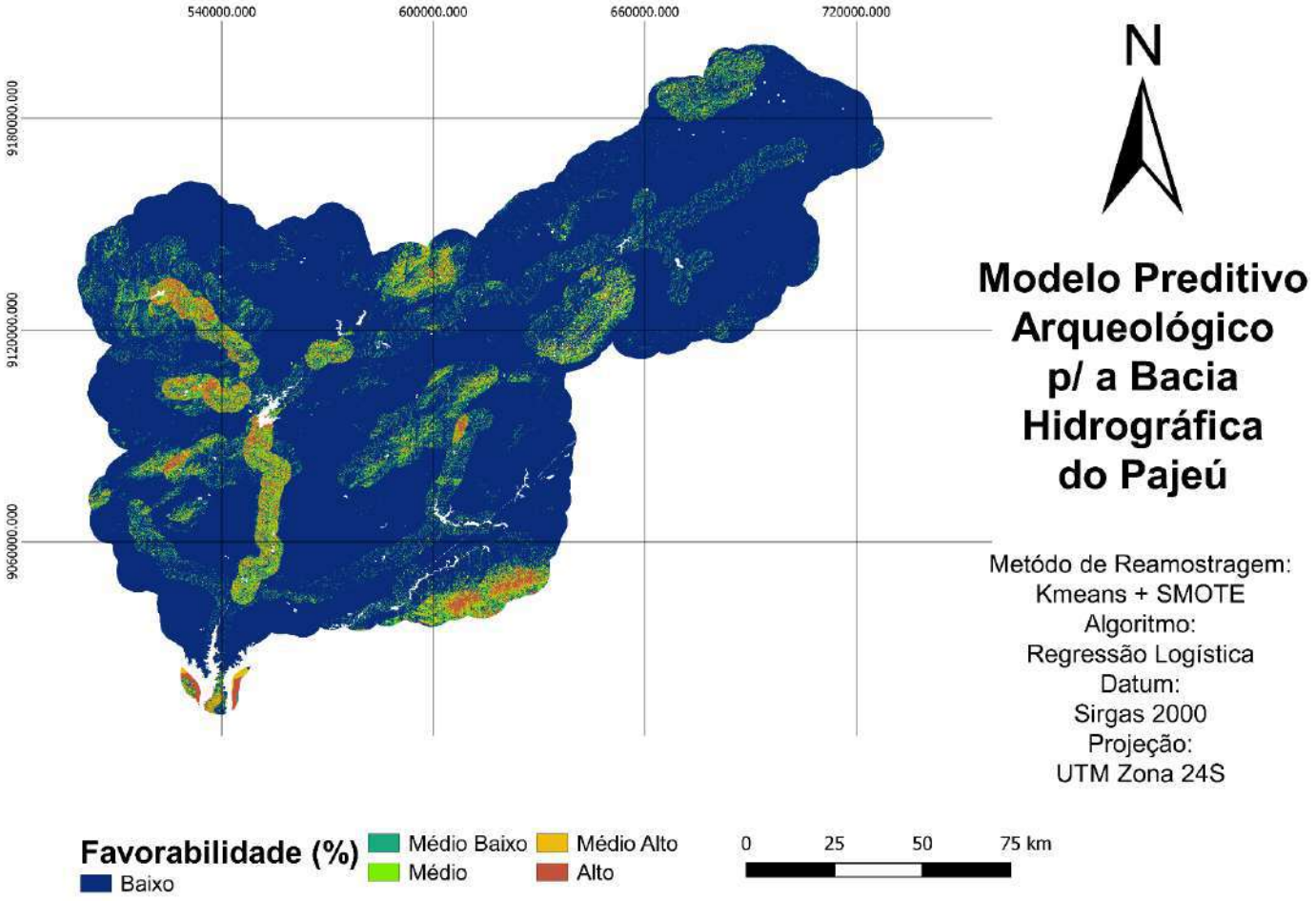
Foram criados ao todo 11 modelos distintos, dos quais três eram regressões logísticas, cada um com seu método de reamostragem específico (*Adasyn*, *Smote*, *KMeans Smote*); três florestas aleatórias, também relacionadas cada um ao método de reamostragem; um modelo de FAB; um modelo de *bagging* balanceado; por fim, três modelos *ensemble*, que unem cada modelo referente ao método de reamostragem usado, com a FAB e o *bagging* balanceado, formando um novo modelo misto.

Figura 25: MPA Inicial com o método de reamostragem ADASYN (Regressão Logísitica)



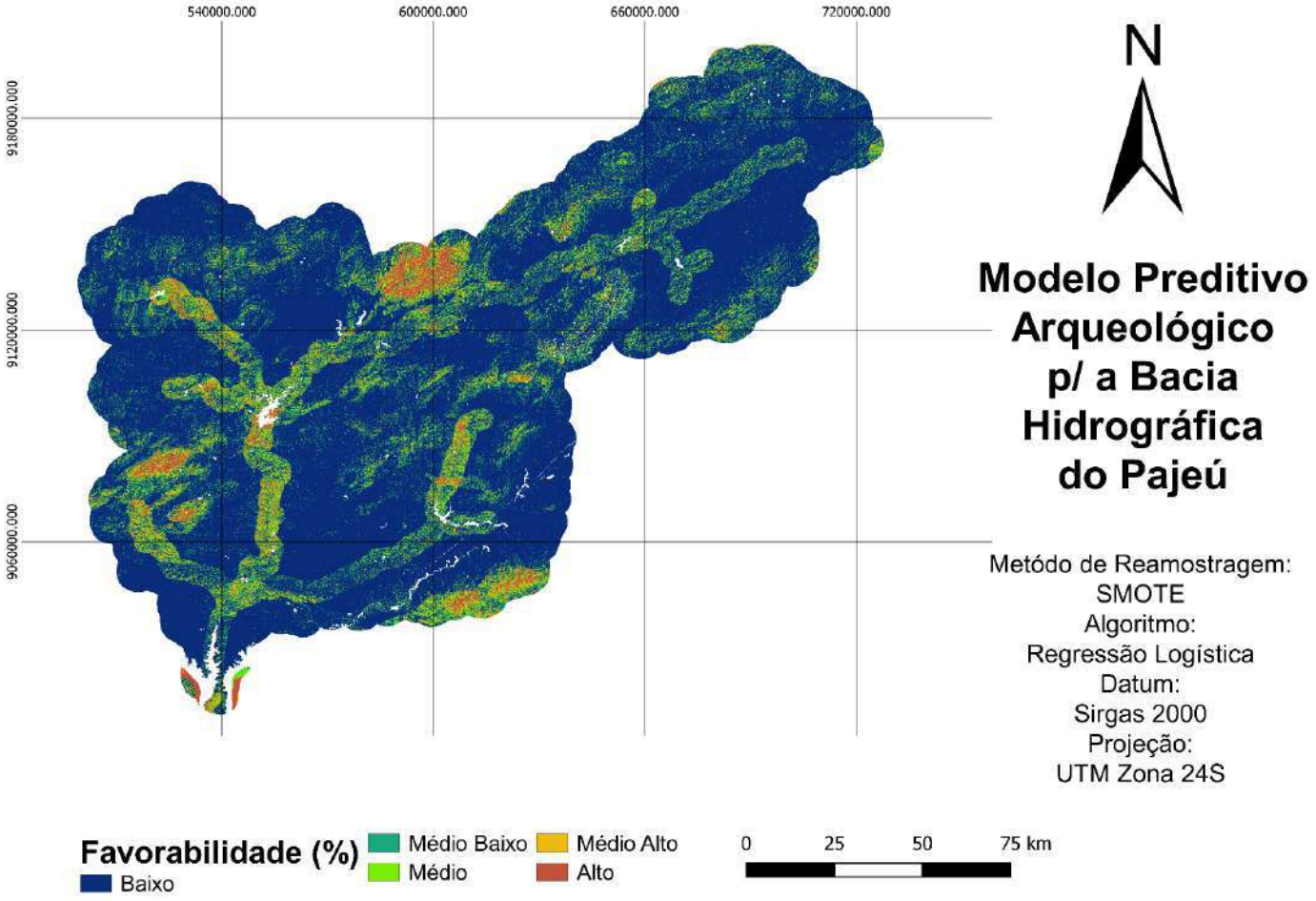
Fonte: Elaborado pelo próprio autor (2021)

Figura 26: MPA Inicial com o método de reamostragem Kmeans + SMOTE (Regressão Logísitica)



Fonte: Elaborado pelo próprio autor (2021)

Figura 27: MPA Inicial com o método de reamostragem SMOTE (Regressão Logísitica)



Fonte: Elaborado pelo próprio autor (2021)

Observando os mapas acima (**Figuras 25, 26 e 27**), é possível perceber que independentemente do método de reamostragem escolhido, a configuração das áreas de alta, média e baixa favorabilidade, permanecem bem similares, com diferenças pontuais. Locais como aquela grande mancha, no centro norte do Pajeú, onde está localizado a Serra da Baixa Verde, ou ainda as áreas da Serra do Umã, Serra do Arapuá, no centro oeste do mapa; ou ainda a Serra Negra, no extremo sudeste da região, são ambientes que aparecem em destaque em todos os modelos, como áreas com alta probabilidade de se achar sítios arqueológicos.

É possível observar também os locais que compreendem o leito dos principais rios da Bacia, aparecem como áreas de média favorabilidade, com pequenas concentrações de alta. A presença de locais com média favorabilidade indica que há uma certa dúvida do modelo em colocar determinadas regiões como positivas ou negativas, o que pode ser explicado por uma falta de informações para àquele determinado ambiente, seja do ponto de vista dos sítios arqueológicos ou das áreas de ausência, ou ainda pelo uso de imagens de satélite com resoluções insuficientes para extrair informações mais específicas para a modelagem.

Com relação a avaliação do modelo com base nas métricas trazidas na metodologia, o modelo apresenta, no geral, um bom desempenho, já que o *recall* foi de 1, ou seja, ele acertou todos os sítios arqueológicos com base no conjunto de validação e teste, ao passo que a precisão ficou baixa, mas como informado, isto não é um problema sério aqui, já que o mais importante é observar as áreas positivas. O detalhamento das avaliações pode ser visto nas tabelas abaixo.

Tabela 1: Métricas estatísticas de avaliação para o modelo de regressão logística

Regressão Logísitica			
	Adasyn	**Kmeans Smote**	**Smote**
F2 Score	0.724637681	0.753768844	0.753768844
Precisão	0.344827586	0.379746835	0.379746835
Recall	1	1	1

Fonte: elaborado pelo próprio autor (2021)

Tabela 2: Matriz de confusão regressão lógistica

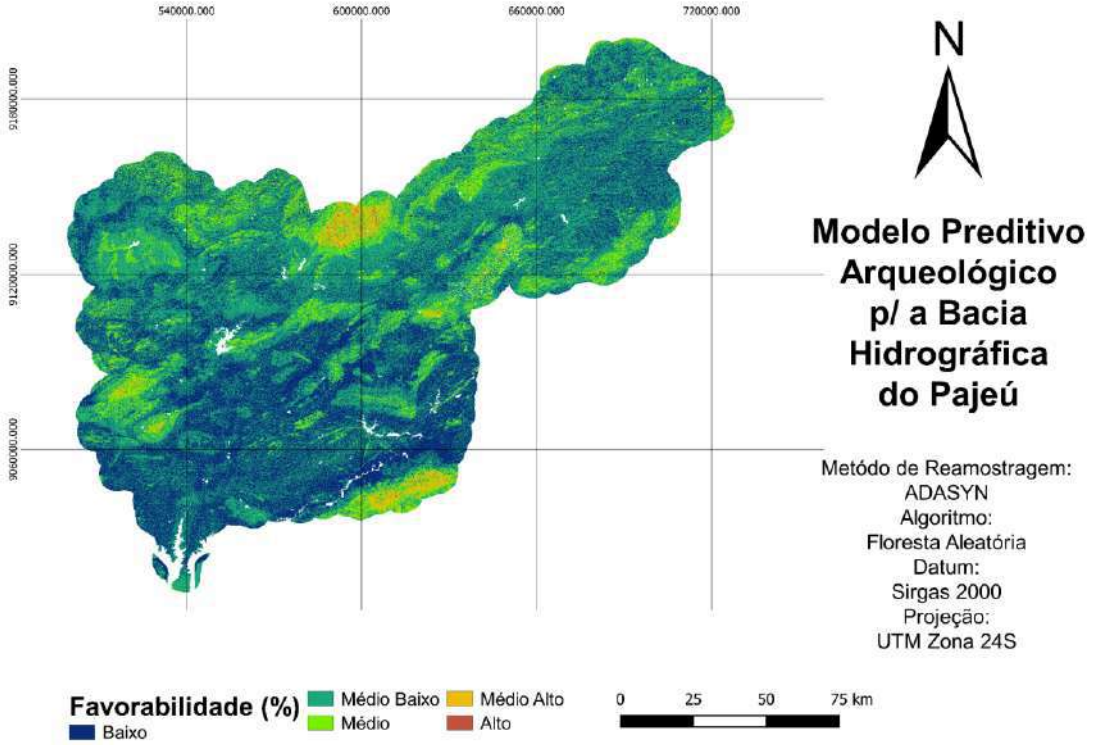
Regressão logística						
	Smote		Adasyn		Kmeans Smote	
	Ausência	Presença	Ausência	Presença	Ausência	Presença
Ausência	290	49	282	57	290	49
Presença	0	30	0	30	0	30

Fonte: elaborado pelo próprio autor (2021)

É possível observar nos valores trazidos nas tabelas acima (**Tabelas 1 e 2**), tanto em relação as métricas como na própria configuração da matriz de confusão, que todos os modelos têm performances muito similares, que não são traduzidas apenas nos mapas gerados. Esses resultados foram bem satisfatórios para uma primeira aplicação dos modelos.

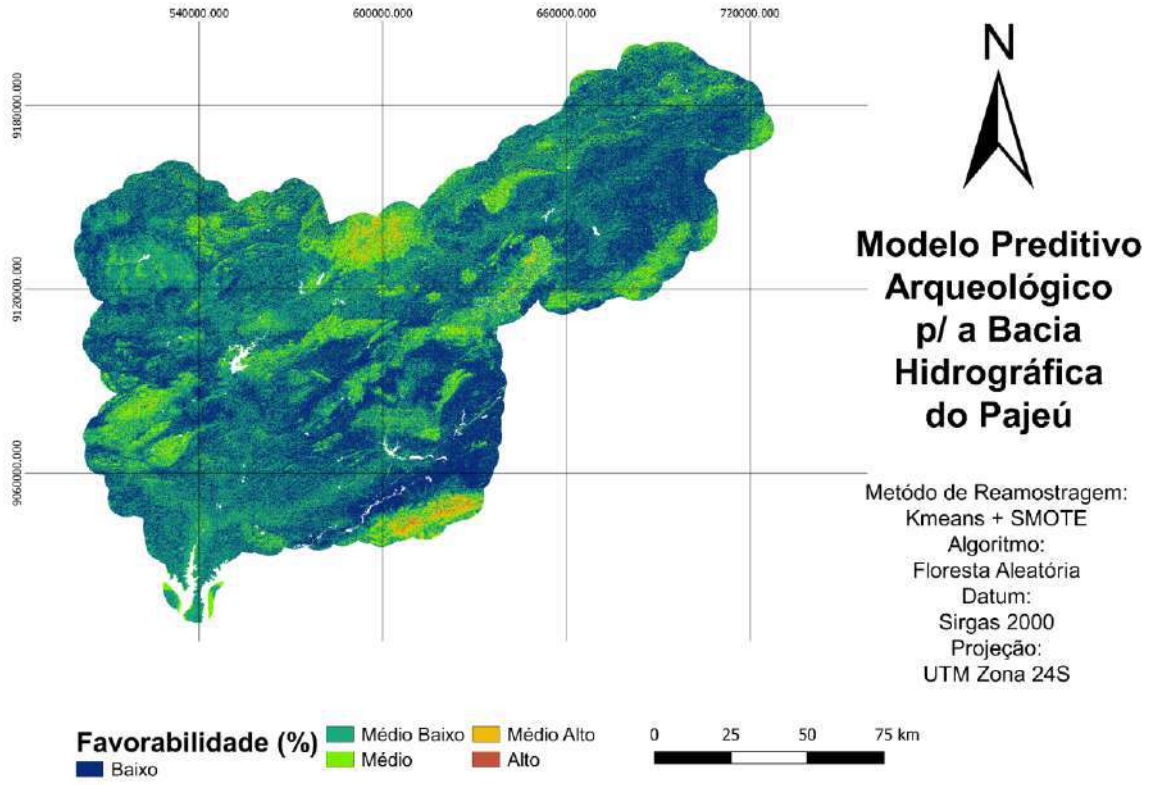
O segundo grupo de modelos que podemos apresentar são os referentes às florestas aleatórias. Assim como a regressão logística, serão feitos um modelo para cada método de reamostragem e os resultados podem ser observados nos mapas abaixo (**Figuras 28, 29 e 30**)

Figura 28: MPA Inicial com o método de reamostragem ADASYN (Floresta Aleatória)



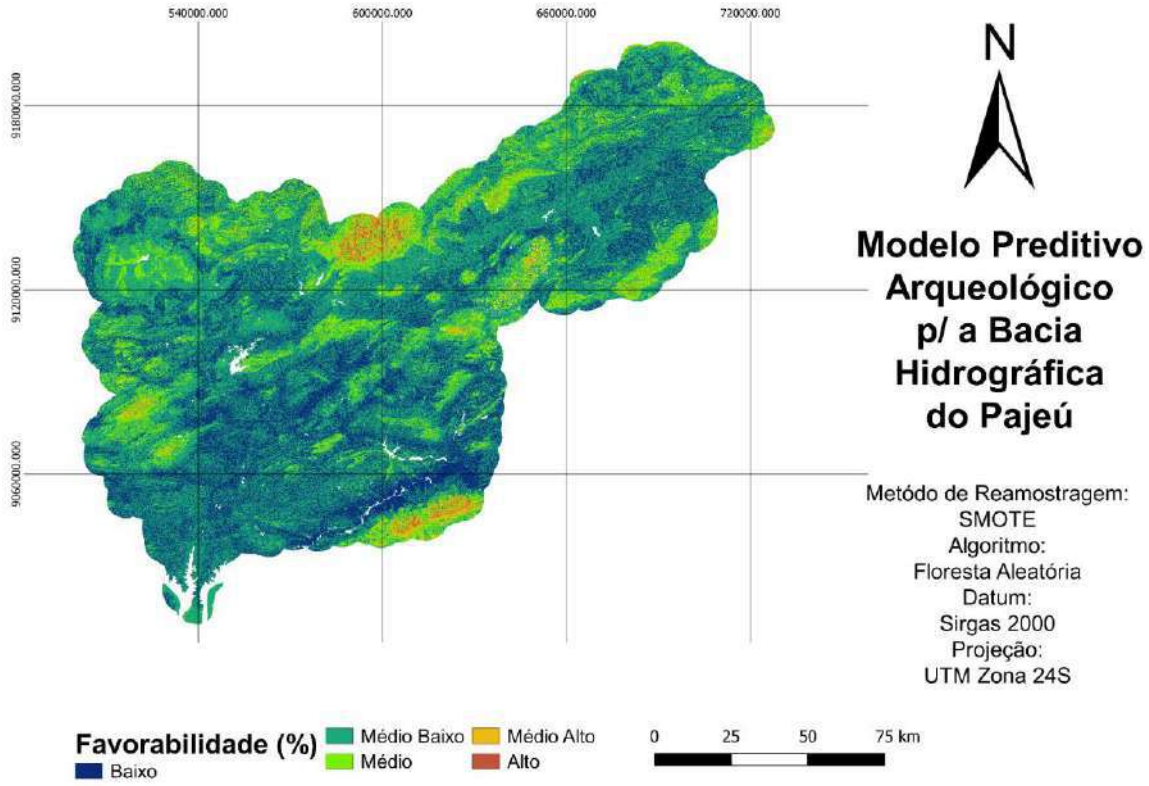
Fonte: Elaborada pelo próprio autor (2021)

Figura 29: MPA Inicial com o método de reamostragem Kmeans + SMOTE (Floresta Aleatória)



Fonte: Elaborado pelo próprio autor (2021)

Figura 30: MPA Inicial com o método de reamostragem SMOTE (Floresta Aleatória)



Fonte: Elaborado pelo próprio autor (2021)

Como é possível observar nos três mapas acima, as áreas da Serra da Baixa Verde a norte e da Serra Negra, ao sul, ainda continuam sendo os locais com maior concentração de píxels com grande probabilidade de se achar sítios arqueológicos. No modelo também é possível observar uma grande quantidade de manchas de média probabilidade, o que reforça a ideia de falta de informações suficientes a serem extraídas pelos modelos.

Um outro ponto que chama atenção ao olhar os mapas acima, é com relação ao modelo que utiliza o *KMeans Smote* (**Figura 27**) como método de reamostragem. Fica nítido pelas imagens que o modelo identificou pouquíssimos locais de alta probabilidade, quando comparado aos outros dois exemplos, indicando uma maior rigorosidade na avaliação das áreas.

Avaliando estatisticamente cada um dos modelos, é possível observar que os maiores valores de Recall e F2 foi do exemplo que utilizou o *KMeans Smote* como método de reamostragem. No geral, os valores apresentados nas tabelas abaixo (**Tabelas 3 e 4**) e a configuração dos mapas, indicam que de fato há muitas áreas de dúvidas, que ficaram no limiar entre presença e ausência, fazendo o modelo ter resultados satisfatórios, mas que ao serem observados friamente com os mapas, denotam falhas e insuficiência de dados.

Tabela 3: Métricas estatísticas de avaliação para o modelo de floresta aleatória

Floresta Aleatória			
	Adasyn	**Kmeans Smote**	**Smote**
F2 Score	0.856164384	0.927152318	0.878378378
Precisão	0.961538462	0.903225806	0.928571429
Recall	0.833333333	0.933333333	0.866666667

Fonte: Elaborado pelo próprio autor (2021)

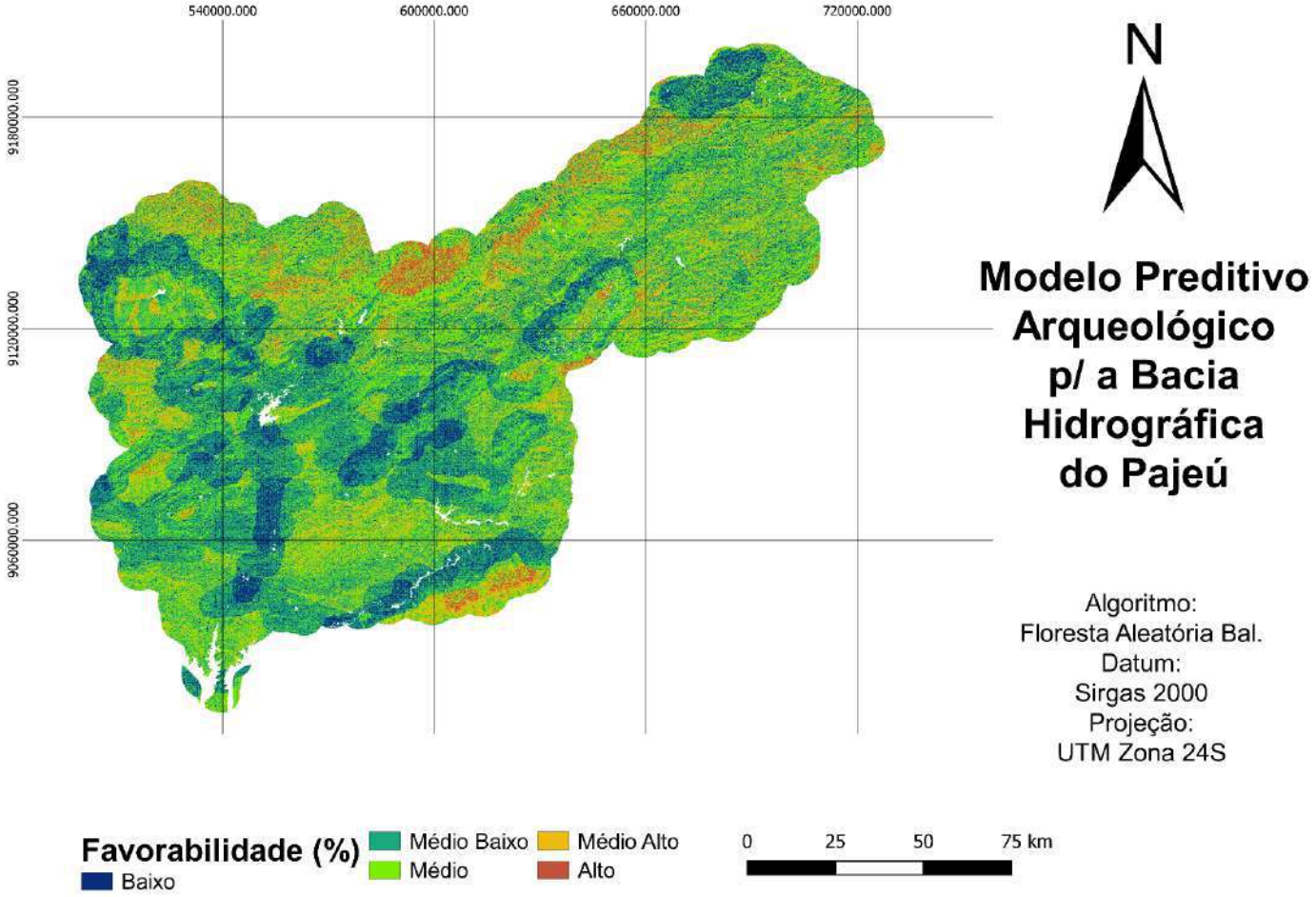
Tabela 4: Matriz de confusão floresta aleatória

Floresta Aleatória						
	Smote		Adasyn		Kmeans Smote	
	Ausência	Presença	Ausência	Presença	Ausência	Presença
Ausência	334	5	337	2	336	3
Presença	6	24	3	27	3	27

Fonte: Elaborado pelo próprio autor (2021)

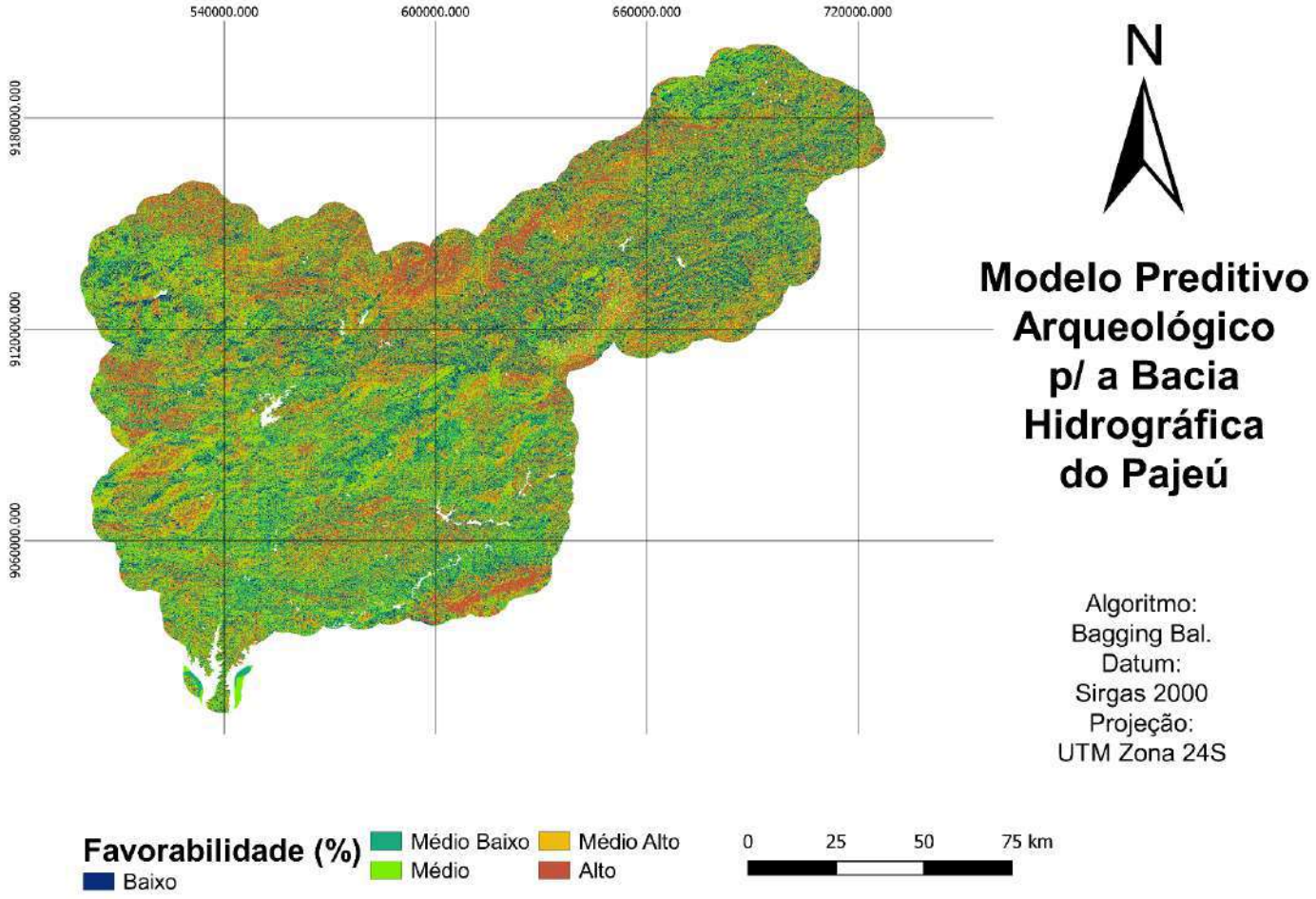
Os modelos que utilizam de algoritmos já prontos para lidar com dados desbalanceados também foram empregados nessa etapa inicial de testagem. A configuração dos mapas tanto de FAB quanto do *bagging* balanceado apresentaram semelhanças como os exemplos anteriores, porém, como é possível ver abaixo (**Figuras 31 e 32**), sobretudo no segundo mapa, há uma generalização bem significativa das áreas de alta probabilidade.

Figura 31: MPA Inicial (Floresta Aleatória Balanceada)



Fonte: Elaborado pelo próprio autor (2021)

Figura 32: MPA Inicial (*Bagging* Balanceado)



Fonte: Elaborado pelo próprio autor (2021)

Como é possível observar, no primeiro mapa, referente ao algoritmo FAB, é possível observar uma distribuição muito similar ao modelo de floresta aleatória tradicional, onde há uma grande mancha de áreas de média probabilidade, que só reforçam a situação de pouca informação nessa fase inicial de aplicação do modelo, ao passo que também é perceptível que os locais como a Serra Negra, Serra da Baixa Verde, Serra do Umã e Serra do Arapuá, continuam a se destacar. Além disso nas localidades à leste da Serra da Baixa Verde, há também uma grande dispersão de machas de alta possibilidade.

Quando os olhos se viram para o mapa do modelo do *bagging* balanceado, a configuração ainda tem esses pontos já observados nos outros modelos também em destaque, porém a generalização dos locais de alta probabilidade é muito maior. Isso pode ser explicado pelo uso de um outro tipo de reamostragem, já que ao contrário de todas as outras, esse algoritmo aplica um sub ajuste, retirando informações da classe majoritária, para igualar a minoritária, dando uma dimensão mais equiparável aos locais com possibilidade de se encontrar registros rupestres.

Os resultados das métricas foi bem positivo, sobretudo no que diz respeito ao recall e ao F2, com valores acima dos 80%, como é possível ver na tabela abaixo (**Tabela 5**), o que indica que todos os modelos têm um bom desempenho a nível estatístico. É preciso salientar, porém, que há falta de informações nestes modelos, pois ainda há grandes áreas com média probabilidade (no caso do FAB) ou há uma dispersão exagerada de locais com alta probabilidade (no caso do *bagging* balanceado), isto acaba inflando o resultado final.

Tabela 5: Métricas estatísticas de avaliação para os modelos de floresta aleatória balanceada e *bagging* balanceado

	***Bagging* Balanceado**	**Floresta Aleatória Balanceada (FAB)**
F2 Score	0.853658537	0.953929539
Precisão	0.353658537	0.638297872
Recall	0.966666667	1

Fonte: Elaborado pelo próprio autor (2021)

Tabela 6:Matriz de confusão FAB e *Bagging* Balanceado

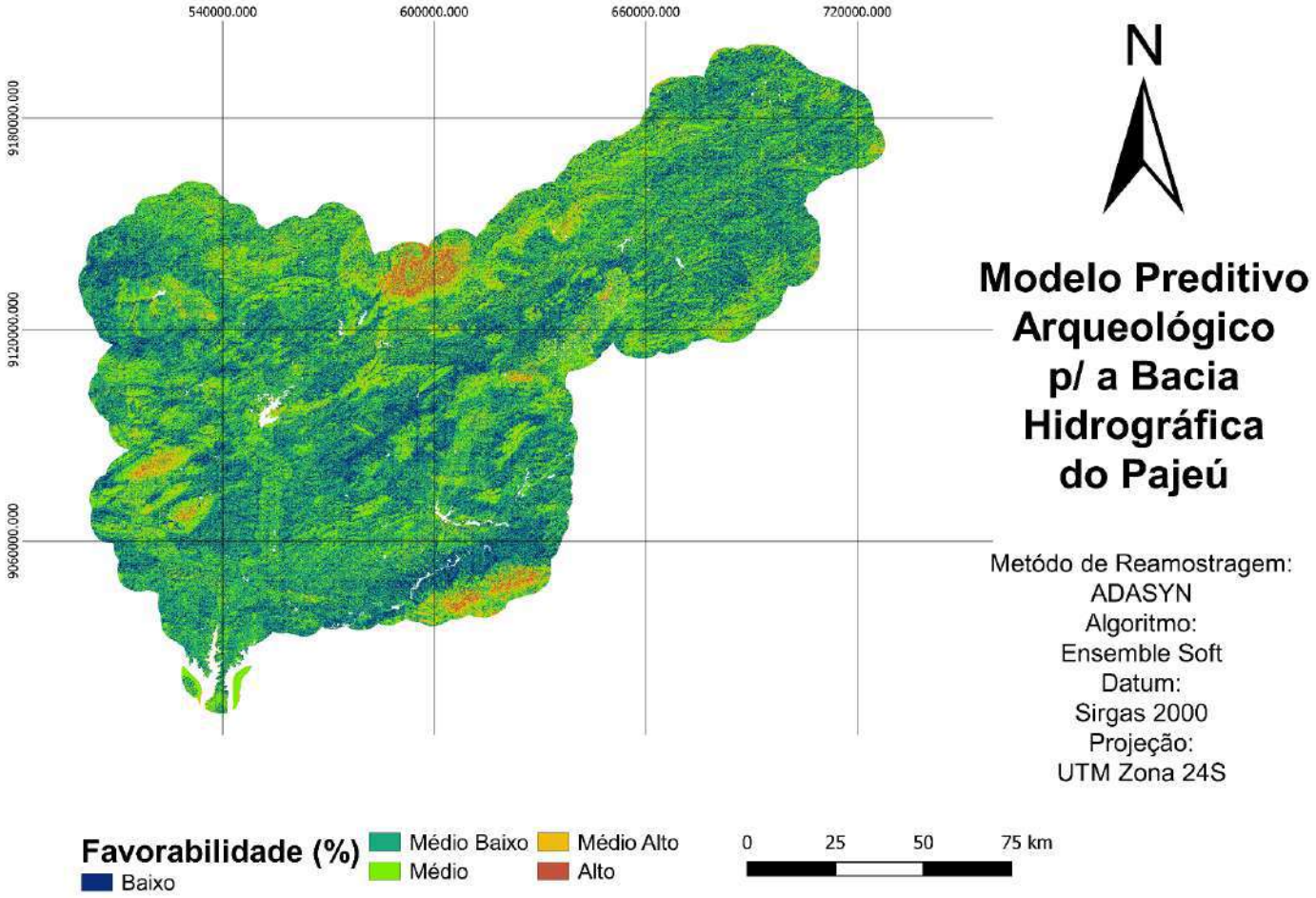
	Bagging Bal.		Floresta Aleatória Bal. (FAB)	
	Ausência	Presença	Ausência	Presença
Ausência	285	54	322	17
Presença	1	29	0	30

Fonte: Elaborado pelo próprio autor (2021)

Os últimos modelos gerados nessa fase inicial foram os *Ensemble*, eles são misturas de todos os outros modelos aqui apresentados. Os resultados dos três modelos referentes aos diferentes métodos de reamostragem, foram bem similares e, novamente, evidenciam os as Serras do Umã, do Arapuá, da Baixa Verde e Negra, como os locais com mais alta probabilidade de se encontrar sítios arqueológicos com grafismos rupestres.

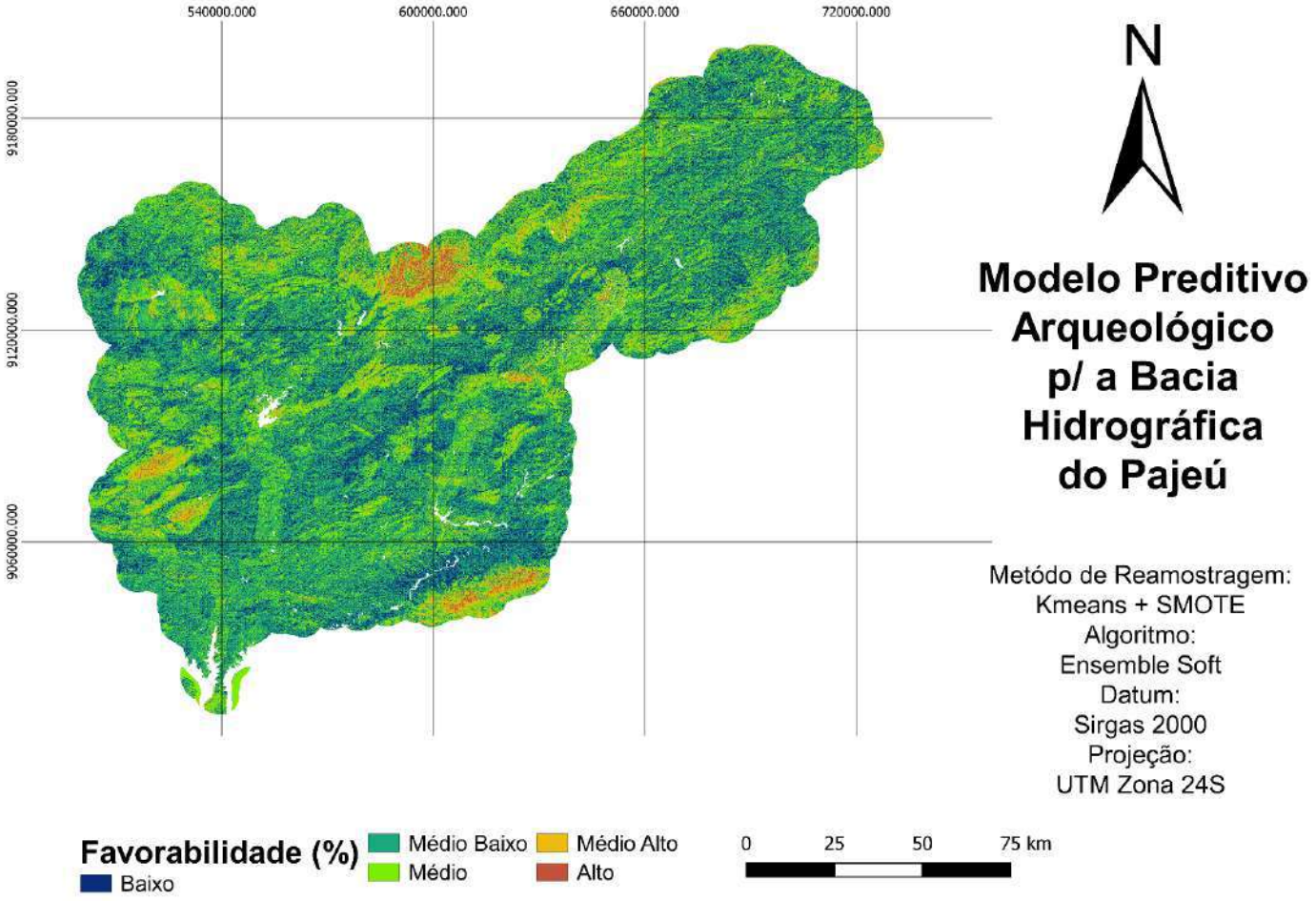
Nos mapas abaixo (**Figuras 33, 34 e 35**), é possível ver algumas outras áreas pontuais com probabilidade alta de presença de registros rupestres também e além disso há também muitos locais com média probabilidade, mas que não são tão latentes como em outros modelos, o que mostra que esse algoritmo, misturando vários outros, obteve um comportamento mais satisfatório, o que pode ser corroborado pelos números abaixo (**Tabelas 7 e 8**).

Figura 33: MPA Inicial com o método de reamostragem ADASYN (Ensemble)



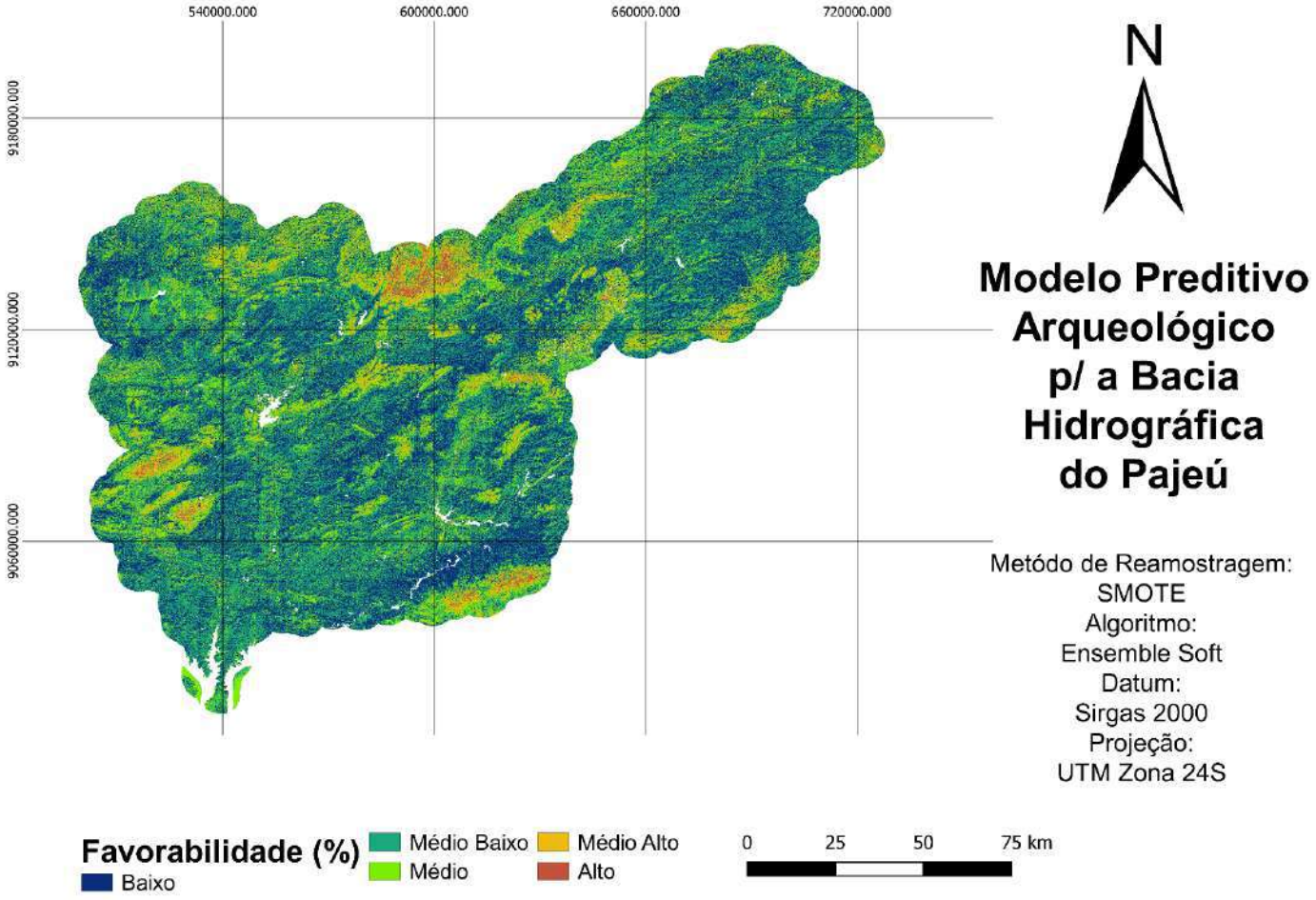
Fonte: Elaborado pelo próprio autor (2021)

Figura 34: MPA Inicial com o método de reamostragem Kmeans + SMOTE (Ensemble)



Fonte: Elaborado pelo próprio autor (2021)

Figura 35: MPA Inicial com o método de reamostragem SMOTE (Ensemble)



Fonte: Elaborado pelo próprio autor (2021)

Tabela 7: Métricas estatísticas de avaliação para o modelo ensemble

Ensemble			
	Adasyn	**Kmeans Smote**	**Smote**
F2 Score	0.877192982	0.867052023	0.872093023
Precisão	0.588235294	0.566037736	0.576923077
Recall	1	1	1

Fonte: Elaborado pelo próprio autor (2021)

Tabela 8: Matriz de confusão modelo Ensemble

Ensemble						
Smote			**Adasyn**		**Kmeans Smote**	
	Ausência	**Presença**	**Ausência**	**Presença**	**Ausência**	**Presença**
Ausência	317	22	318	21	316	23
Presença	0	30	0	30	0	30

Fonte: Elaborado pelo próprio autor (2021)

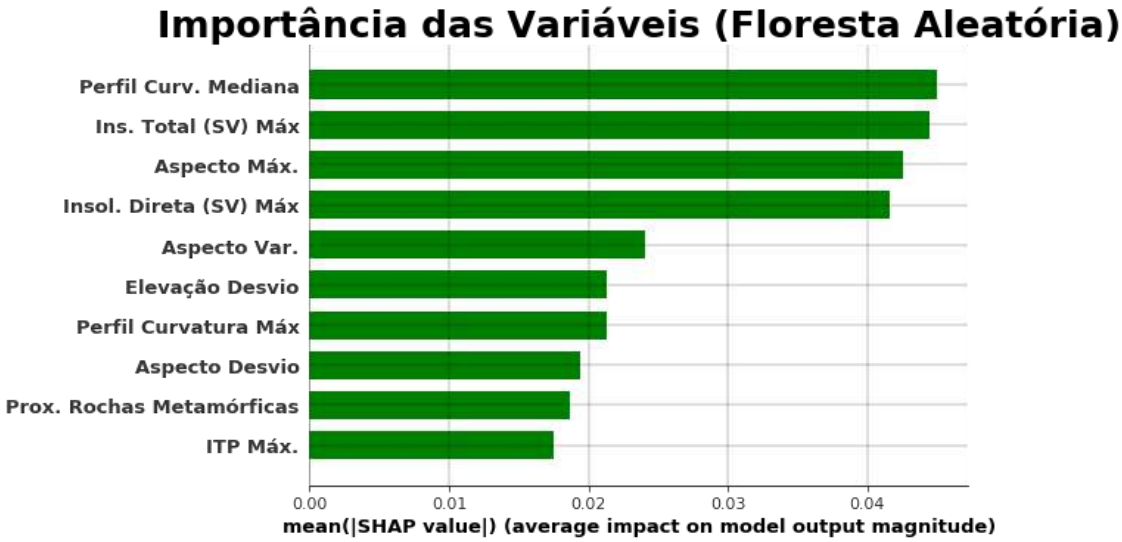
.5.2 A importância das variáveis nos modelos iniciais

Avaliar as métricas por si, não vai nos informar muito a respeito da distribuição espacial e na compreensão dos padrões ambientais e sociais que estão relacionados com os sítios arqueológicos de grafismos rupestres da Região do Pajeú. Um ponto importante a se analisar são os pesos de cada variável que participa do cálculo da modelagem

Para os modelos iniciais aqui expostos foi realizada a extração da importância das variáveis de três algoritmos diferentes, dois deles a partir dos modelos de RL e Floresta Aleatória, ambos utilizando o *SMOTE* como método de reamostragem. Este foi escolhido devido, primeiramente, ao fato de o referido método ter sido o mais significativo estatisticamente, mas também à similaridade entre os mapas das diferentes reamostragens de um mesmo algoritmo, que não apresentam grandes alterações como pode ser visto nos gráficos abaixo (**Figuras 36, 37 e 38**).

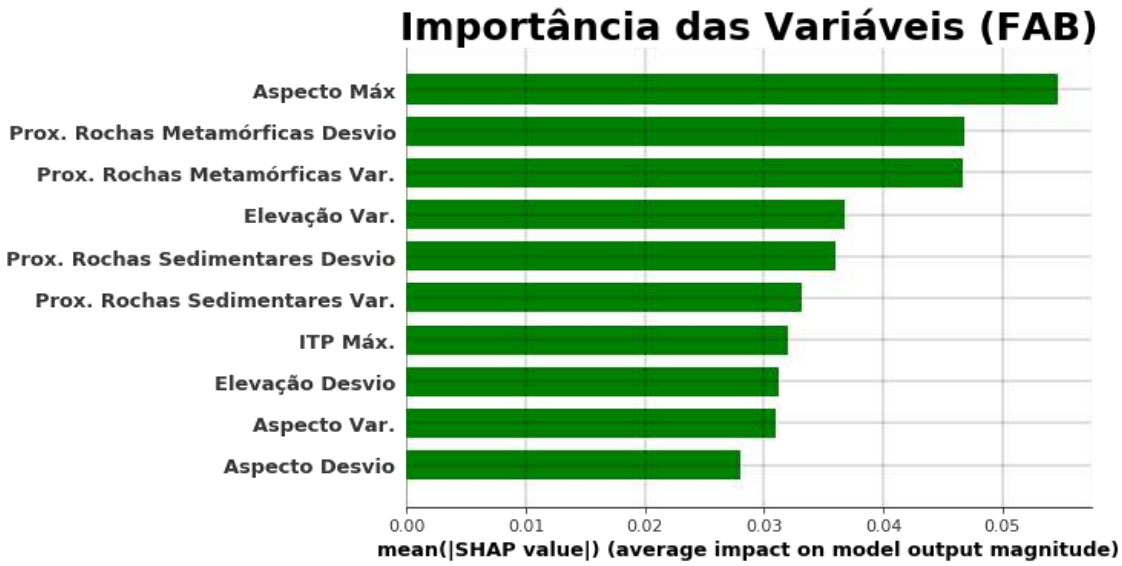
O outro gráfico de importância de variáveis foi extraído do modelo de FAB, o qual usa um método de reamostragem diferente dos demais, além de ter tido uma configuração que apresenta uma série de diferenças quando olhamos os mapas gerados. É importante pontuar ainda que todos os gráficos foram gerados utilizado o *SHAP*, mencionado no capítulo anterior.

Figura 36: Histograma de importâncias das variáveis referentes ao modelo de floresta aleatória



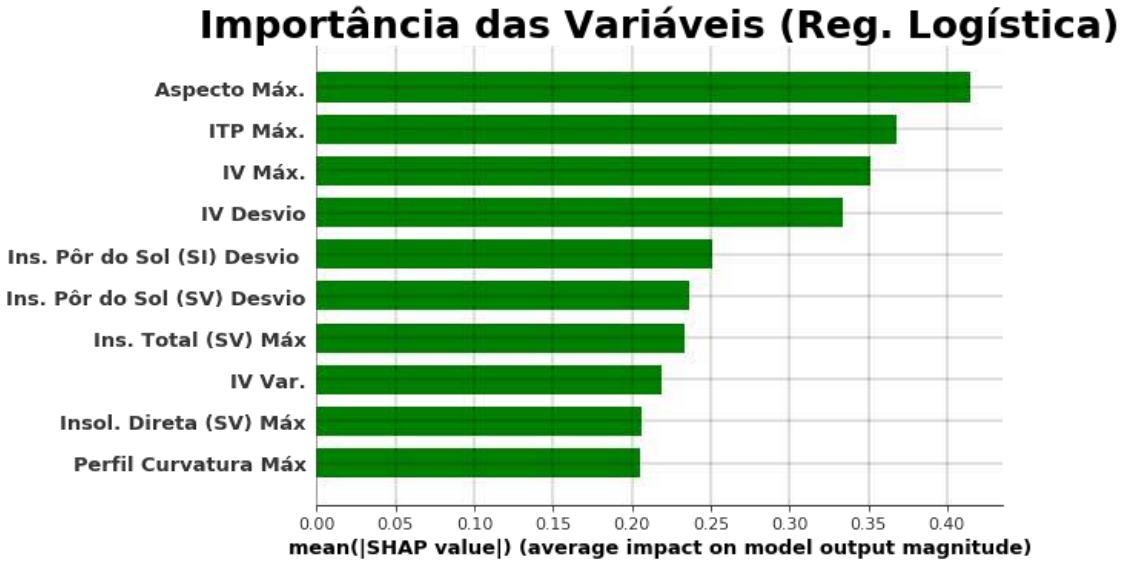
Fonte: Elaborado pelo próprio autor (2021)

Figura 37: Histograma de importâncias das variáveis referentes ao modelo de FAB



Fonte: Elaborado pelo próprio autor (2021)

Figura 38: Histograma de importâncias das variáveis referentes ao modelo de regressão logística e *smote* como método de reamostragem



Fonte: Elaborado pelo próprio autor (2021)

Observando os gráficos acima de maneira geral, percebemos que o “Aspecto Máximo” aparece com destaque nos três diferentes modelos, indicando que a orientação da encosta é, sem dúvida, uma variável de extrema importância ao se buscar por áreas com grafismos rupestres. Olhando mais especificamente para os modelos de RL e Floresta Aleatória temos que variáveis como “ITP Máximo” e “Perfil de Curvatura Máximo” também se destacam.

Analisando diretamente essas variáveis em destaque, observa-se, que dentro de cada área de 100m x 100m, utilizada para calcular a estatística das áreas cinzas, o aspecto ou a direção das encostas é uma variável fundamental na relação com os sítios arqueológicos de grafismos rupestres, o que de certa forma faz sentido, já que dentro dos estudos que analisam a paisagem e os registros parietais, esse é um dos elementos que mais se visualiza .

O ITP e o Perfil de curvatura não são variáveis tão discutidas em trabalhos da área diretamente, porém o primeiro está relacionado diretamente com a elevação do terreno, que sim, é uma variável já referenciada em outras pesquisas no semiárido nordestino, onde temos sítios sobretudo em locais de média encosta . Já o perfil de curvatura, está muito atrelado a formação de áreas abrigadas no terreno. A maioria dos sítios do Pajeú estão em locais assim, portanto faz sentido que esta variável esteja aqui presente.

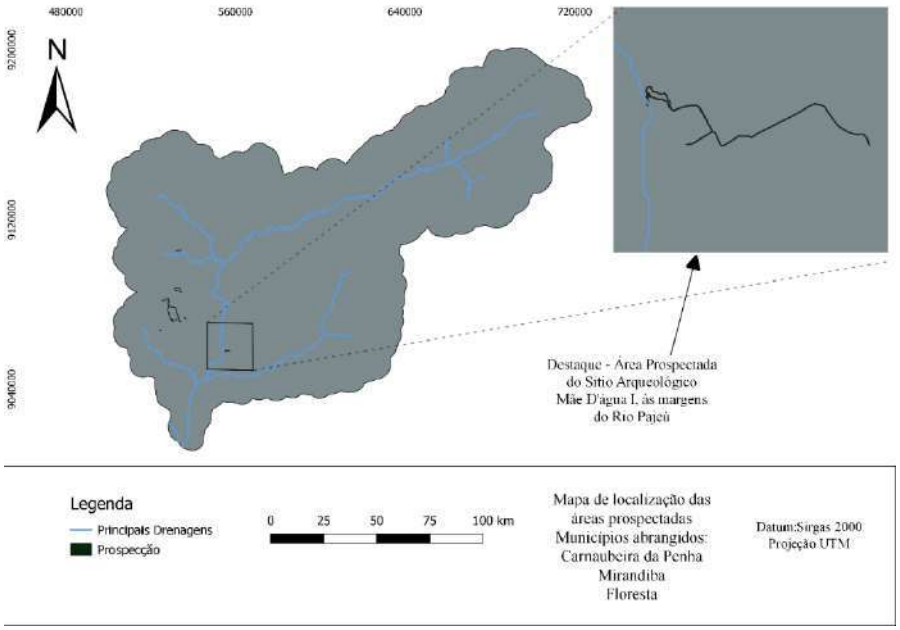
Em resumo, observar a orientação da encosta, aliado ao posicionamento topográfico do local e as curvaturas do relevo, são pontos essenciais para se encontrar vestígios arqueológicos gravados ou pintados nas rochas, de acordo com esses modelos iniciais.

.5.3 O teste em campo

Para testar na prática a performance do modelo foi criado um protocolo para uma primeira visita a campo. Esta foi realizada durante o mês de junho de 2021, portanto ainda sob o auge da crise provocada pelo Covid 19, e, portanto, não pode ser tão extensa.

Foram visitadas ao todo 5 locais diferentes, levando em conta a presença de outros sítios já conhecidos e os próprios dados que os modelos iniciais geraram. No mapa abaixo (**Figura 39**) é possível ver os trechos prospectados durante a campanha. A princípio não foi encontrado nenhum sítio com registo rupestre que ainda não tivesse sido cadastrado, porém foi possível colher muita informação armazenada em um receptor GNSS de navegação.

Figura 39: Áreas prospectadas durante as atividades de campo de junho de 2021



Fonte: Fonte: Elaborado pelo próprio autor (2021)

Além disso, o campo foi frutífero para entender melhor a respeito do contexto local, possibilitando o contato com as comunidades da região. Elas foram fundamentais para a realização do deslocamento em meio a caatinga fechada durante os escassos 5 dias de prospecção.

Apesar de nenhum sítio novo ter sido avistado, foi possível levantar materiais importantes, com os habitantes locais, que indicam que há locais a serem registrados na área. Foram coletadas algumas fotos e vídeos que mostram pinturas ainda não catalogadas por outras equipes (**Figura 40**). Além disso nas margens do Pajeú, em uma das regiões prospectadas ao sul, e onde muitos modelos indicavam um local de baixa probabilidade, foram encontradas mais gravuras, que a princípio não haviam sido catalogadas de maneira precisa.

Figura 40: Sítio não cadastrado com grafismos rupestres na Serra das Flores, Carnaubeira da Penha

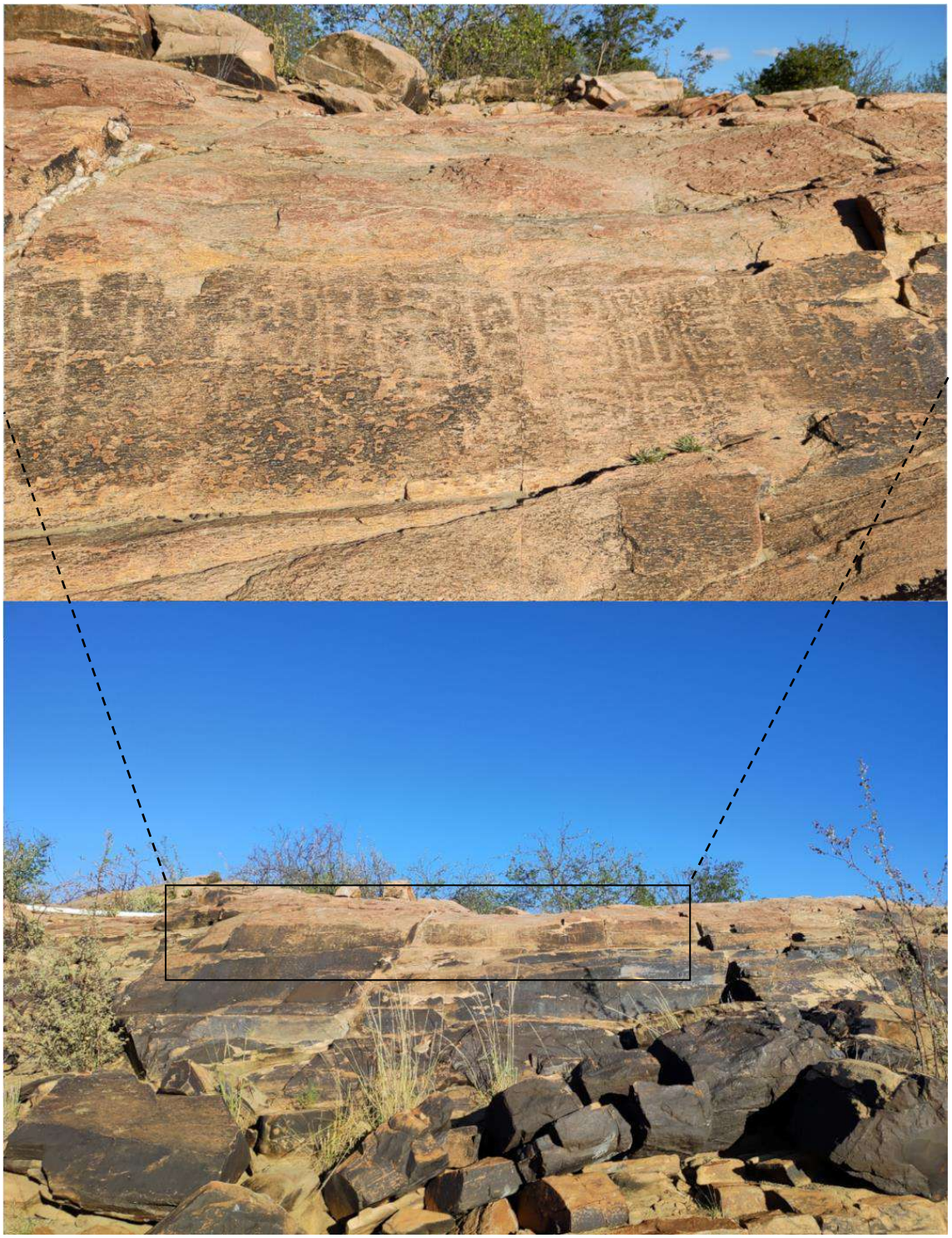
Fonte: Foto tirada por Joaquim Camilo (2012)

Este fato, ocorrido no sítio Mãe D'água I, às margens do Rio Pajeú, mostra uma série de gravuras rupestres presentes nas rochas do leito do rio (**Figura 41**). O IPHAN, dentro do Cadastro Nacional de Sítios Arqueológicos, coloca esta região como um grande complexo de gravuras, com vários quilômetros de extensão. Observando a distribuição das principais gravuras do local, pudemos separar a área em mais dois sítios, como é possível ver no mapa abaixo, de modo a se adequar a configuração do modelo.

Esta situação nos leva a pensar na própria delimitação de sítio arqueológico que tomamos e que é um norteador desta pesquisa. Será que considerar uma área tão vasta como um único sítio arqueológico, é correto? E Dividi-la em diversos sítios? A verdade é que não há um certo nem errado nesta situação, o fato de se ter dividido a área em 3, foi apenas para poder dar mais informações ao modelo, e enquadrar os sítios dentro de uma área de 100m², que foi a

média calculada inicialmente no trabalho. Como as áreas rebaixadas, mais próximas dos principais cursos d'água do Pajeú, mostraram ser as mais difíceis de serem compreendidas pelos algoritmos, toda informação nova nessas porções, será bem-vinda com o intuito de aprimorar os produtos gerados.

Figura 41: Gravuras nas margens do Rio Pajeú



Fonte: Foto tirada pelo próprio autor (2021)

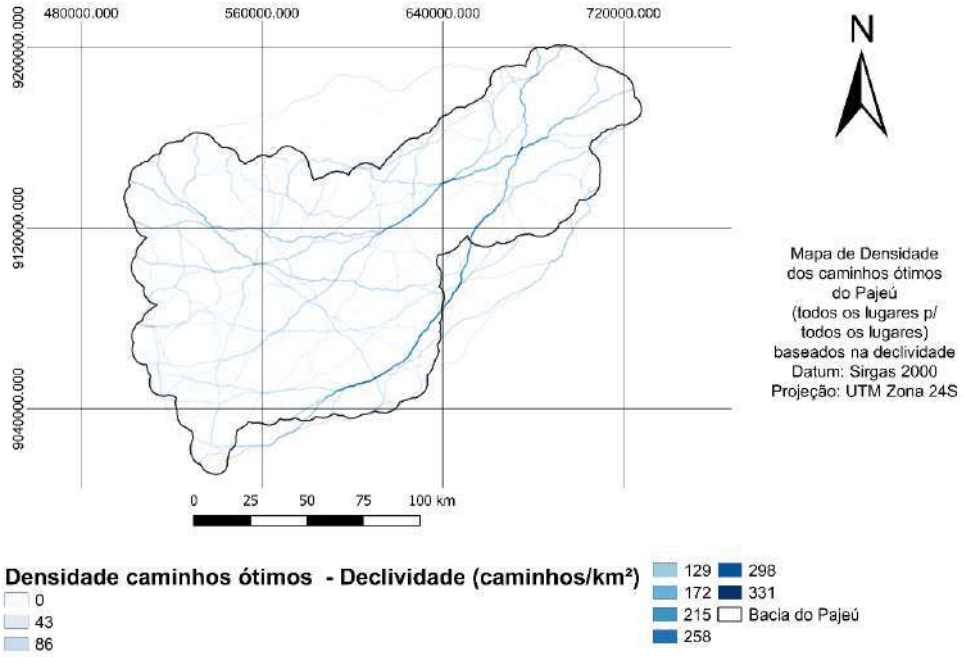
6. MPA_PAJEU_1.0: MELHORANDO OS MODELOS

Após a ida a campo e com a inserção dos novos dados relatados no capítulo anterior, novos modelos foram então executados, utilizando também novas variáveis. Ao todo foram criadas mais 4 colunas à tabela de atributos, e todas elas foram derivadas da acessibilidade. Elas usaram a declividade do terreno e a insolação como superfície de custo para a criação de uma rede de caminhos ótimos, gerando tanto uma variável baseada na densidade de rotas, como também na proximidade dos sítios e não sítios, aos CMC.

6.1 O desenvolvimento dos modelos

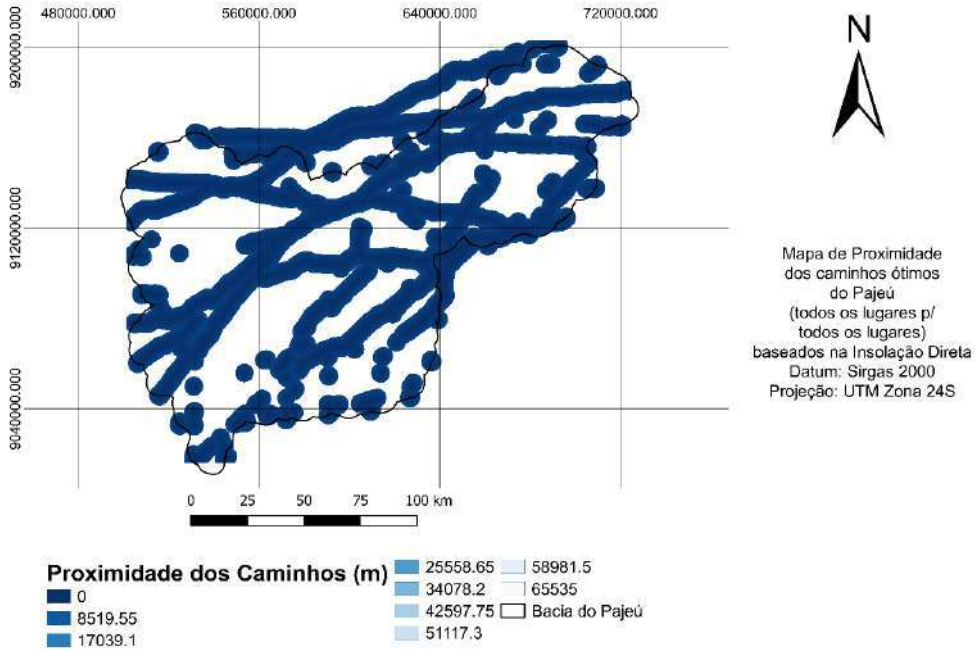
O grande diferencial destes modelos criados na etapa final do trabalho, foi a inserção das variáveis derivadas da acessibilidade. Estas puderam fornecer um novo olhar sobre as informações de cunho socio-ambiental para um MPA e fazer melhor entender a respeito de possíveis áreas de deslocamento dos grupos pré-históricos na região do Pajeú e até de áreas vizinhas do semiárido nordestino. Os produtos gerados podem ser vistos nos mapas abaixo (**Figuras 42 e 43)**:

Figura 42: Mapa de densidade dos caminhos ótimos do Pajeú usando a declividade como superfície de custo.



Fonte: Elaborado pelo próprio autor (2022)

Figura 43: Mapa de Proximidade dos caminhos ótimos do Pajeú usando a insolação como superfície de custo.



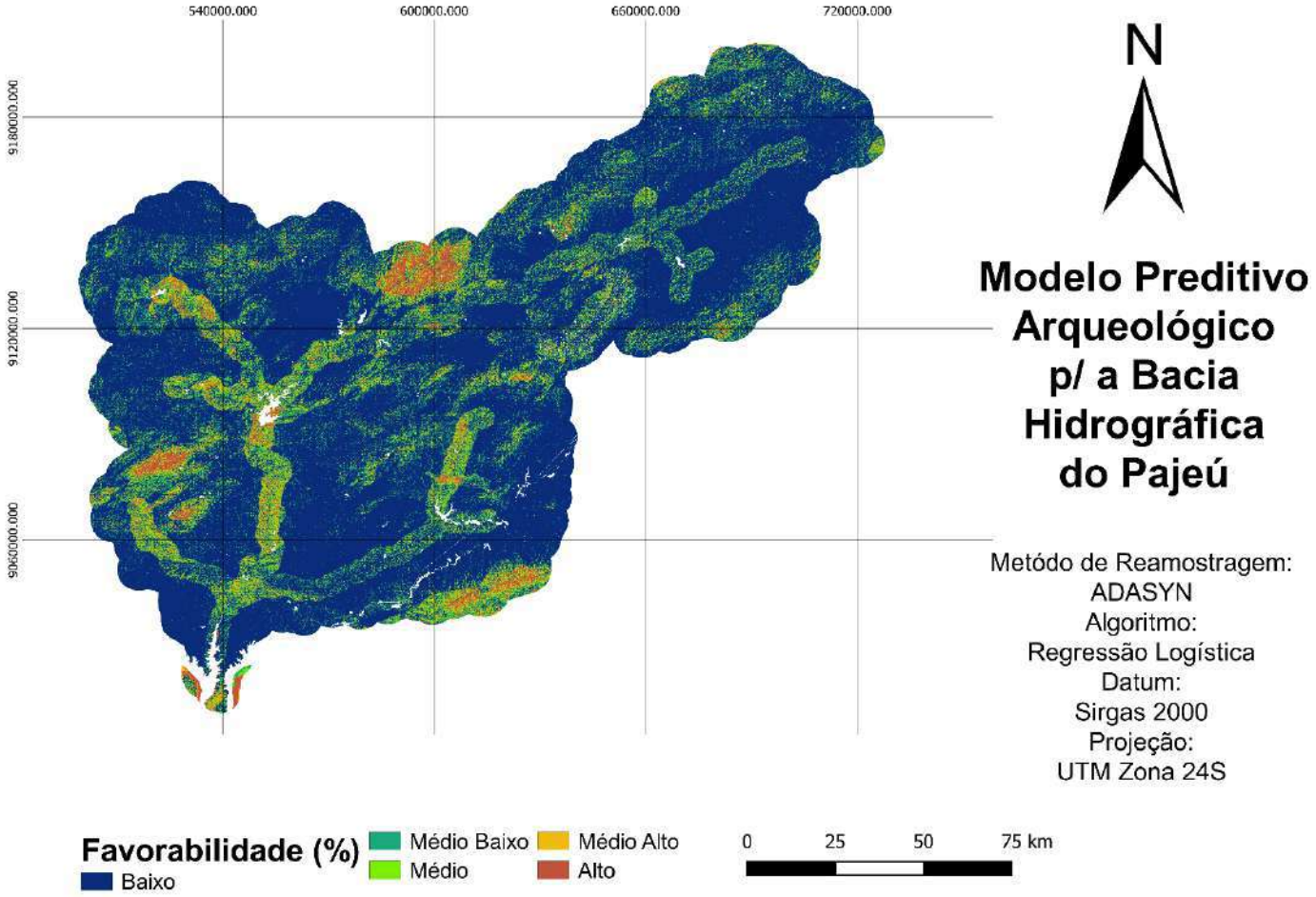
Fonte: Elaborado pelo próprio autor (2022)

O que fica claro analisando os produtos que usaram tanto a declividade como a insolação como superfície de custo, é que há uma diferença bastante significativa na configuração dos caminhos traçados, o que ressalta a importância de se usarem diferentes bases para calcular rotas de menor custo, ainda mais no contexto do semiárido nordestino, que por mais que tenha passado por mudanças climáticas significativas, ainda assim é um local em uma região tropical, em que a insolação pode ter tido influência na escolha dos grupos ao se deslocar.

6.1.2 Os modelos gerados

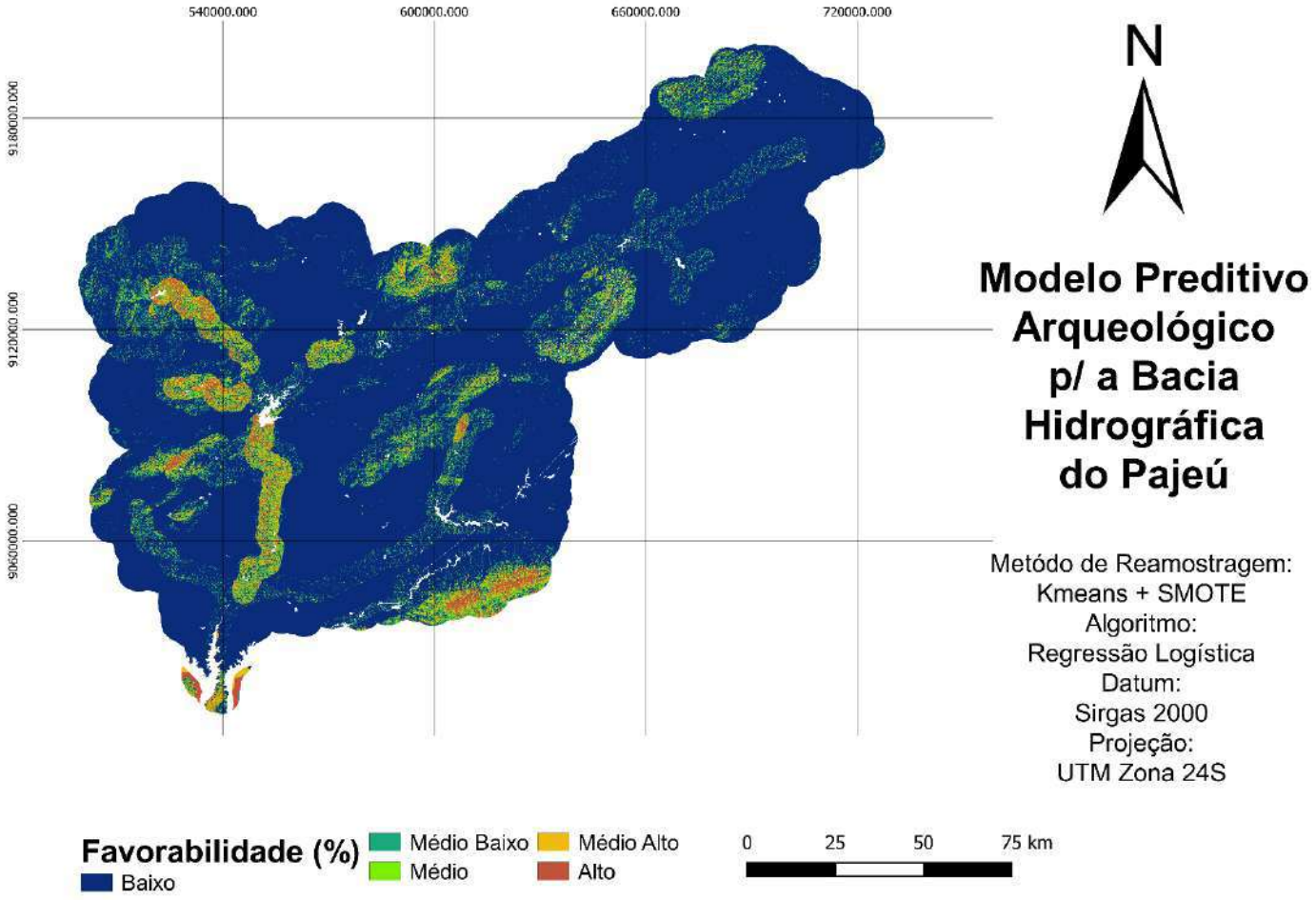
Com a execução destes modelos, foi possível perceber diferenças significativas quando comparamos os resultados pré-campo com estes. De antemão é possível afirmar que, na maioria dos casos, às áreas já destacadas anteriormente (Serra do Arapuá e do Umã, Serra Negra, Serra da Baixa Verde e Serra Vermelha) continuam sendo vistas como locais de alta favorabilidade de encontrar sítios com grafismos rupestres.

Figura 44: MPA_Pajeu_1.0 com o método de reamostragem ADASYN (regressão logística)



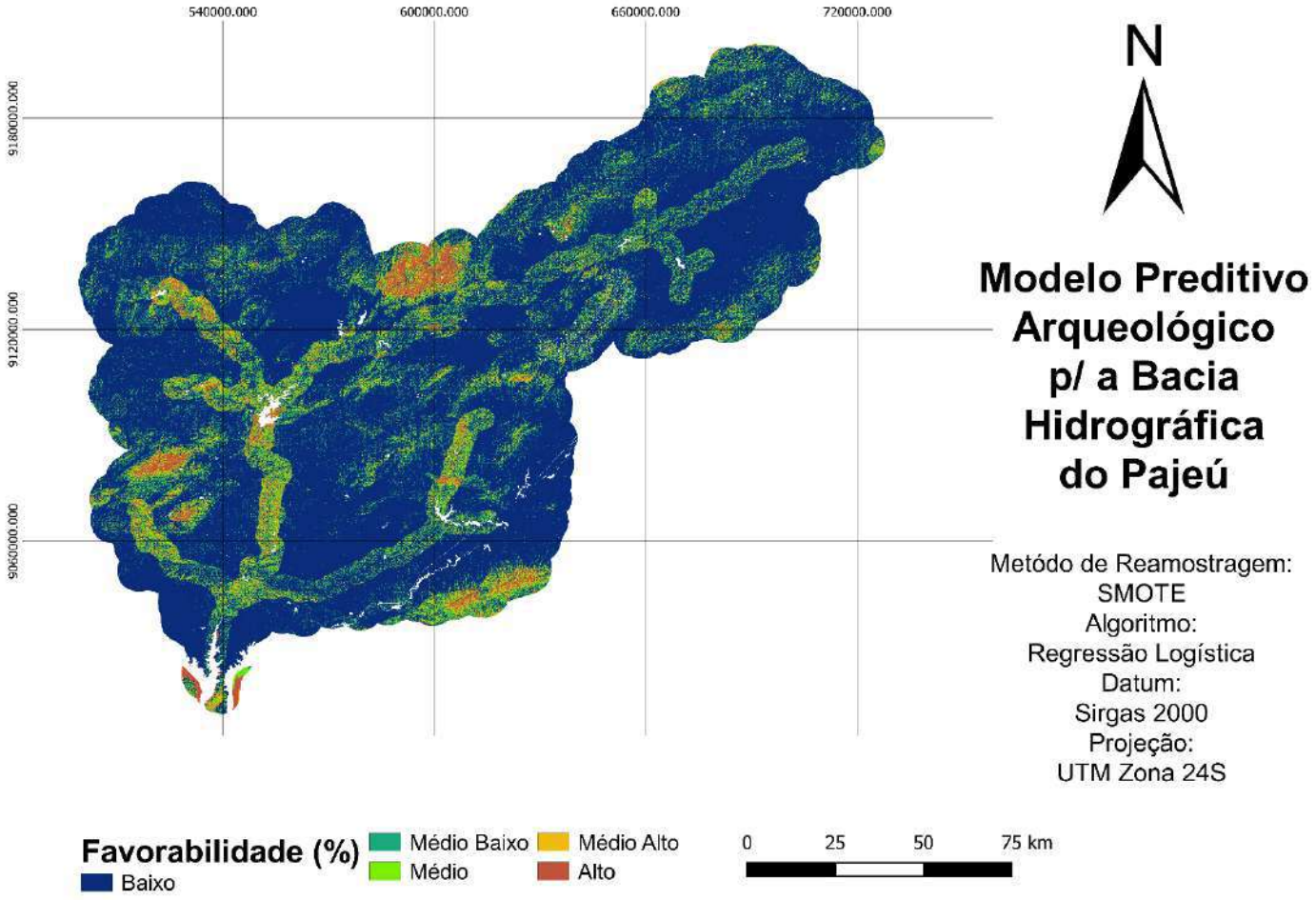
Fonte: Elaborado pelo próprio autor (2022)

Figura 45: MPA_Pajeu_1.0 com o método de reamostragem Kmeans + SMOTE (regressão logística)



Fonte: Elaborado pelo próprio autor (2022)

Figura 46: MPA_Pajeu_1.0 com o método de reamostragem SMOTE (regressão logística)



Fonte: Elaborado pelo próprio autor (2022)

Ao olhar os mapas acima (**Figuras 44, 45 e 46**), é possível perceber que nos modelos de regressão logística, ao serem comparados os diferentes métodos de reamostragem, continua-se vendo configurações bem similares, e como já foi adiantado, às áreas que já vinham sendo percebidas como locais de alta probabilidade, seguem aparecendo. É possível perceber também, que os locais próximos aos leitos das principais drenagens, apresenta pontos favoráveis a serem encontrados sítios arqueológicos de grafismo rupestre.

Ao olharmos os modelos de regressão logística iniciais e finais, é possível perceber que as áreas mais rebaixadas em relação ao entorno, próximos às drenagens de grande ordem, apresentam locais de alta favorabilidade, mas de modo geral, são áreas de média probabilidade e, portanto, locais de dúvida.

Com relação a avaliação estatística, temos resultados similares ao modelo inicial, com uma melhora nos números de precisão, o que já foi discutido anteriormente, não afeta de forma direta na qualidade da modelagem, já que o recall permaneceu alto e é esta métrica que guia os valores positivos de fato.

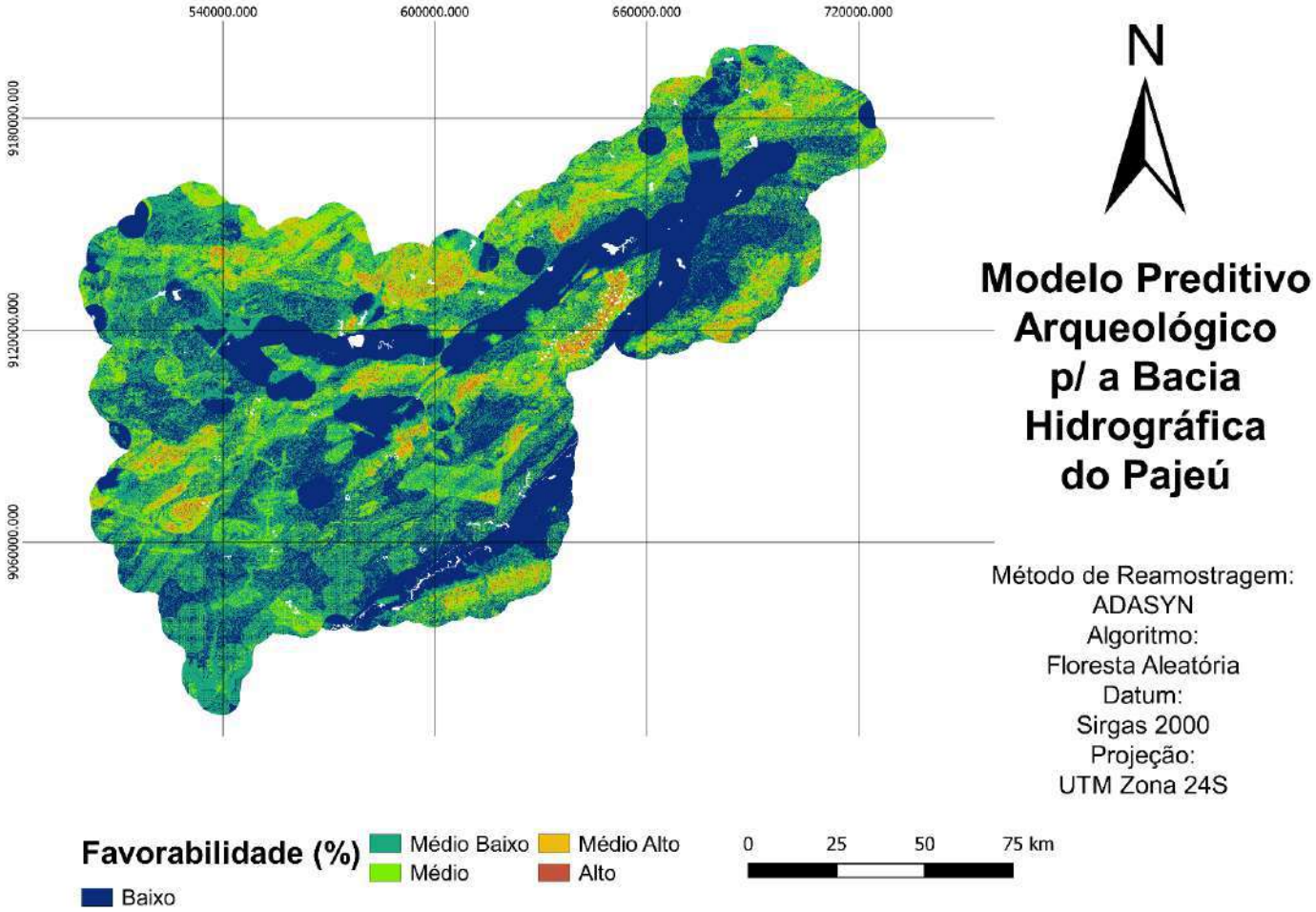
Tabela 9: Métricas estatísticas de avaliação para o modelo final de regressão logísitca

Regressão Logística			
	Adasyn	Kmeans Smote	Smote
F2 Score	0.753	0.764	0.793
Precisão	0.440	0.458	0.469
Recall	0.917	0.917	0.958

Fonte: Elaborado pelo próprio autor (2022)

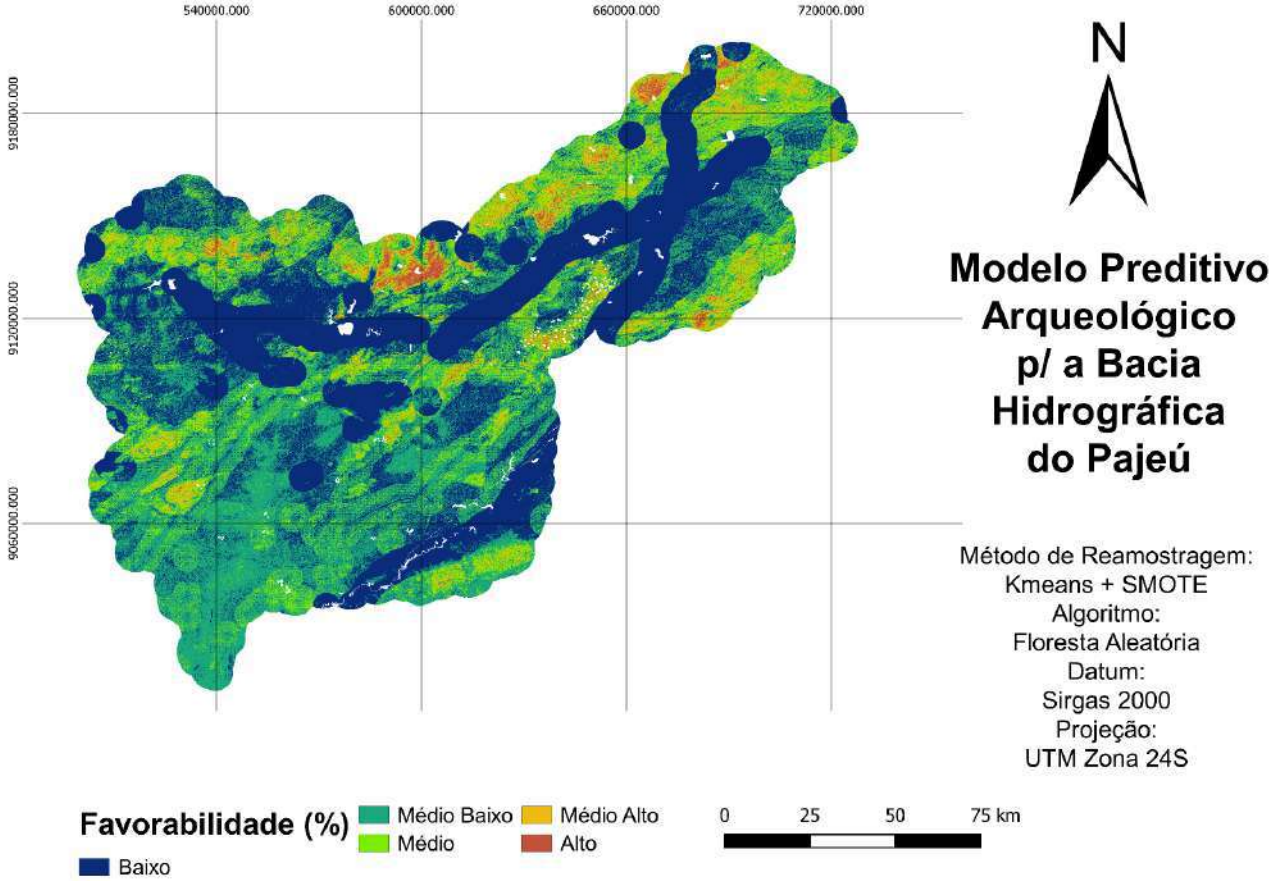
O modelo de Floresta Aleatória continua apresentando uma série de áreas de média probabilidade, igualmente aos modelos iniciais gerados, o que reforça a tese de que há falta de dados, mesmo com a inserção das novas informações coletadas em campo durante o mês de junho de 2021 e as novas variáveis derivadas da acessibilidade. Os mapas podem ser vistos nas figuras abaixo (**Figuras 47, 48 e 49**)

Figura 47: MPA_Pajeu_1.0 com o método de reamostragem ADASYN (floresta aleatória)



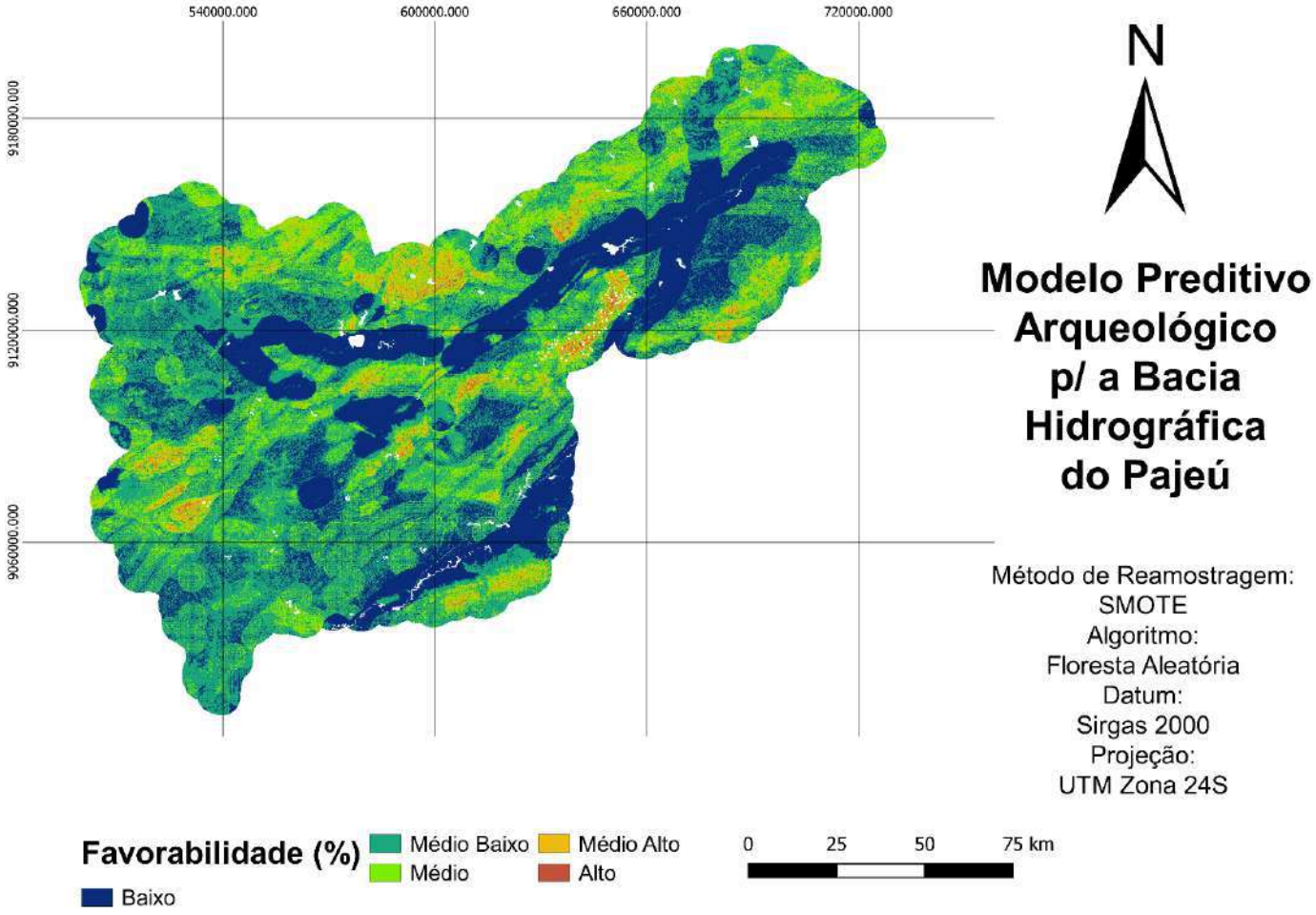
Fonte: Elaborado pelo próprio autor (2022)

Figura 48: MPA_Pajeu_1.0 com o método de reamostragem Kmeans + SMOTE (floresta aleatória)



Fonte: Elaborado pelo próprio autor (2022)

Figura 49: MPA_Pajeu_1.0 com o método de reamostragem SMOTE (floresta aleatória)



Fonte: Elaborado pelo próprio autor (2022)

Olhando para os mapas acima, é possível ver que às áreas das Serras do Umã, Arapuá, Negra, Vermelha e da Baixa Verde, aparecem como locais de alta probabilidade de se encontrar sítios, assim como já vinha sendo mostrado em modelos anteriores. A grande diferença aqui fica por conta da grande faixa de área com baixa favorabilidade de se achar sítios arqueológicos, relacionadas justamente aos CMC referentes a declividade como superfície de custo. Esta configuração mostra que as principais rotas que focam no quesito gasto de energia, como é o caso da declividade, não influenciam no achado de novos sítios arqueológicos, o que corrobora a utilização de outros aspectos ambientais que podem ter contribuído no deslocamento e escolha dos artistas para a realização ou não de grafismos rupestres.

Com relação a avaliação estatística dos modelos o resultado apresentado abaixo (**Tabela 10**) mostra que tanto a precisão, como o *recall*, apresentaram altos valores, mas olhando para os mapas acima, é possível perceber que as áreas de média probabilidade ainda continuam dominantes no mapa, indicando ainda a falta de informações suficientes em determinadas áreas do Pajeú.

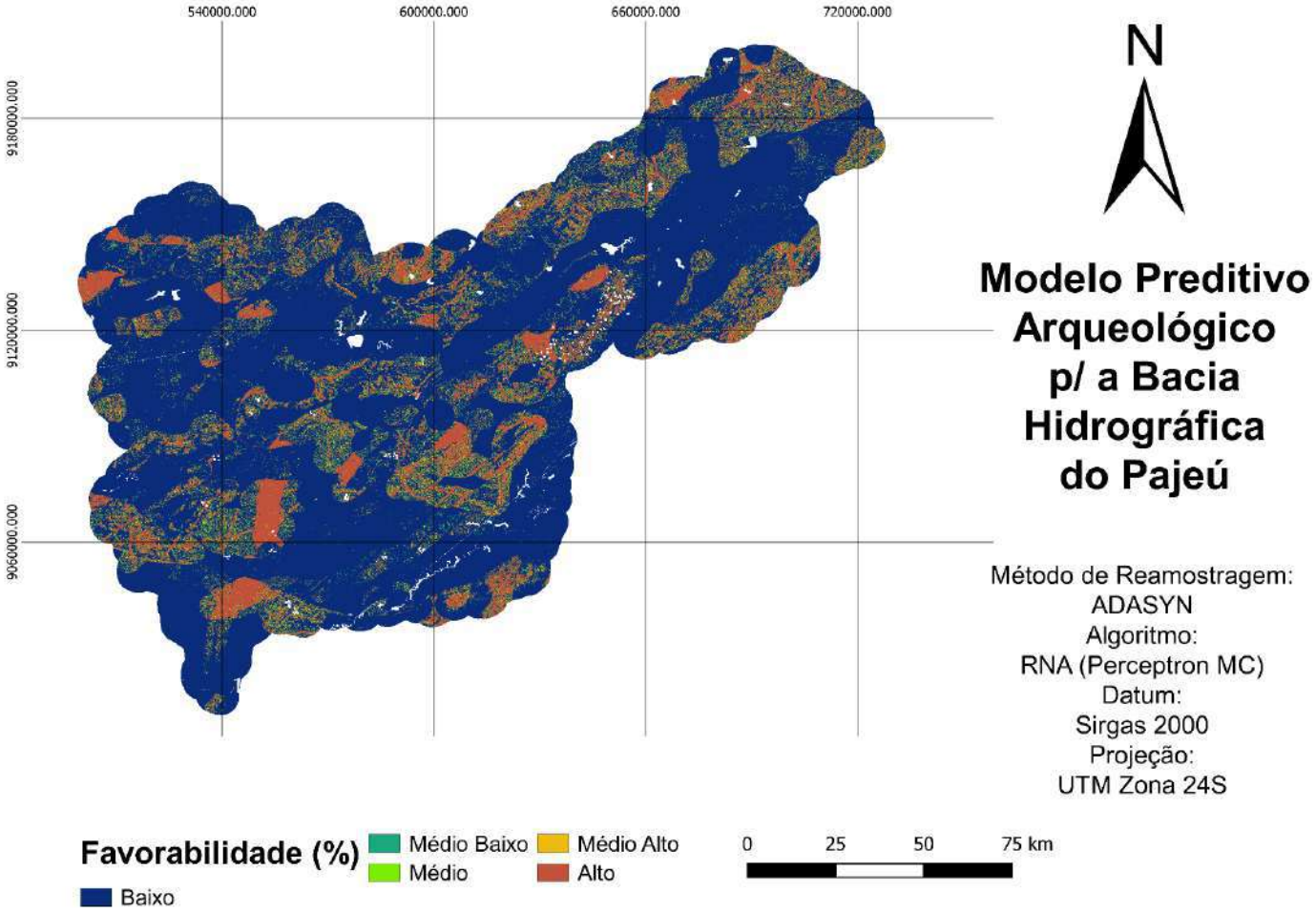
Tabela 10: Métricas estatísticas de avaliação para o modelo final de floresta aleatória

Floresta Aleatória			
	Adasyn	Kmeans Smote	Smote
F2 Score	0.882	0.882	0.890
Precisão	0.913	0.913	0.955
Recall	0.875	0.875	0.875

Fonte: Elaborado pelo próprio autor (2022)

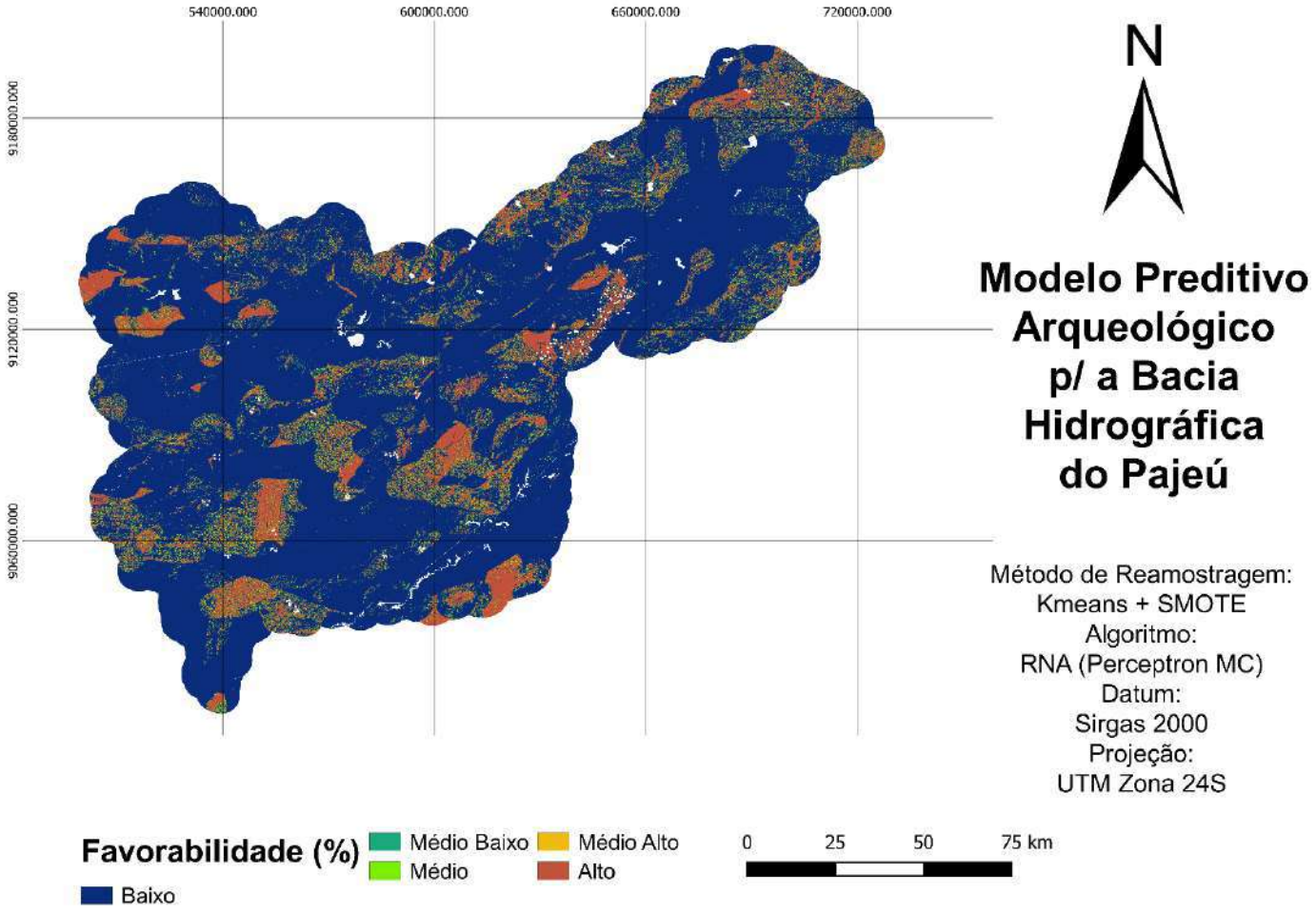
Como já adiantado anteriormente no início do capítulo, durante a construção dos novos modelos pós-campo, foi utilizada RNA (perceptron multicamadas) para a construção de novos modelos em associação aos métodos de reamostragem. Os modelos deste tipo usam uma estrutura matemática similar à da regressão logística e a configuração dos mapas abaixo (**Figuras 50, 51 e 52**), mostram uma similaridade entre os produtos gerados pelos algoritmos em questão.

Figura 50: MPA_Pajeu_1.0 com o método de reamostragem ADASYN (RNA Perceptron)



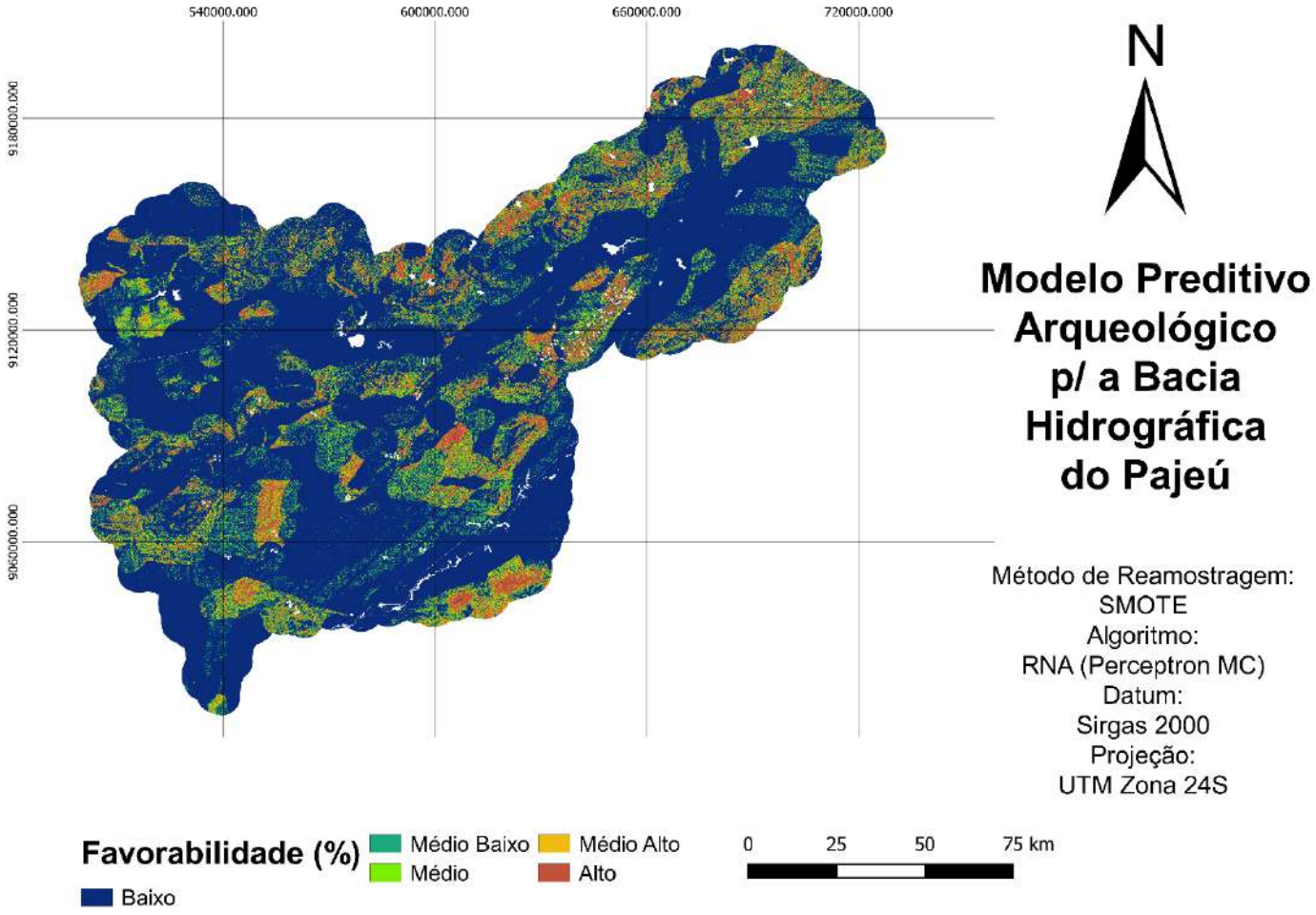
Fonte: Elaborado pelo próprio autor (2022)

Figura 51: MPA_Pajeu_1.0 com o método de reamostragem Kmeans SMOTE (RNA Perceptron)



Fonte: Elaborado pelo próprio autor (2022)

Figura 52: MPA_Pajeu_1.0 com o método de reamostragem SMOTE (RNA Perceptron)



Fonte: Elaborado pelo próprio autor (2022)

Olhando os mapas acima, é possível perceber que as áreas próximas ao leito das principais drenagens aparecem novamente como destaque, assim como nos modelos de regressão logística. Os locais já relatados como destaques anteriormente aparecem nos modelos acima, porém a área de abrangência dos setores de maior probabilidade tem uma densidade maior. Por outro lado, se olharmos para as médias probabilidades, estas aparecem com maior ênfase no modelo que usa o *SMOTE* como método de reamostragem e mesmo assim não aparece de forma tão intensa como nos outros mapas apresentados.

Quando olhamos os números dentro da avaliação estatística dos modelos de RNA (**Tabela 11**), vemos que, ao comparar com os de regressão logística, temos aqui valores de precisão bem mais altos e um recall seguindo basicamente o mesmo padrão. Estes números em união aos mapas acima, mostram que o uso da RNA pode ser um grande aliado na construção de MPA para se encontrar sítios com registros rupestres.

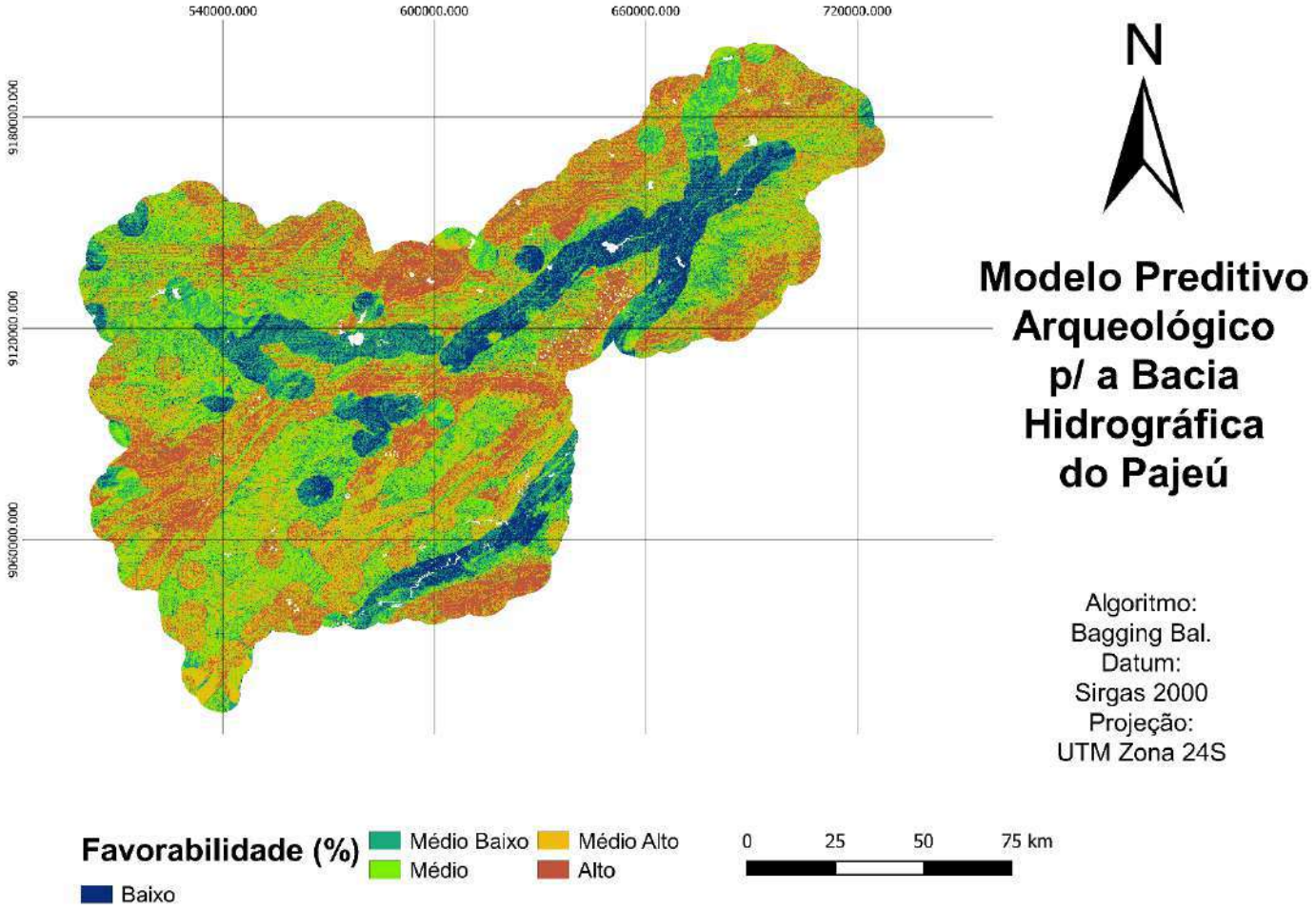
Tabela 11: Métricas estatísticas de avaliação para o modelo final de RNA Perceptron multicamadas

RNA (Perceptron Multicamadas)			
	Adasyn	Kmeans Smote	Smote
F2 Score	0.875	0.847	0.885
Precisão	0.875	0.750	0.676
Recall	0.875	0.875	0.958

Fonte: Elaborado pelo próprio autor (2022)

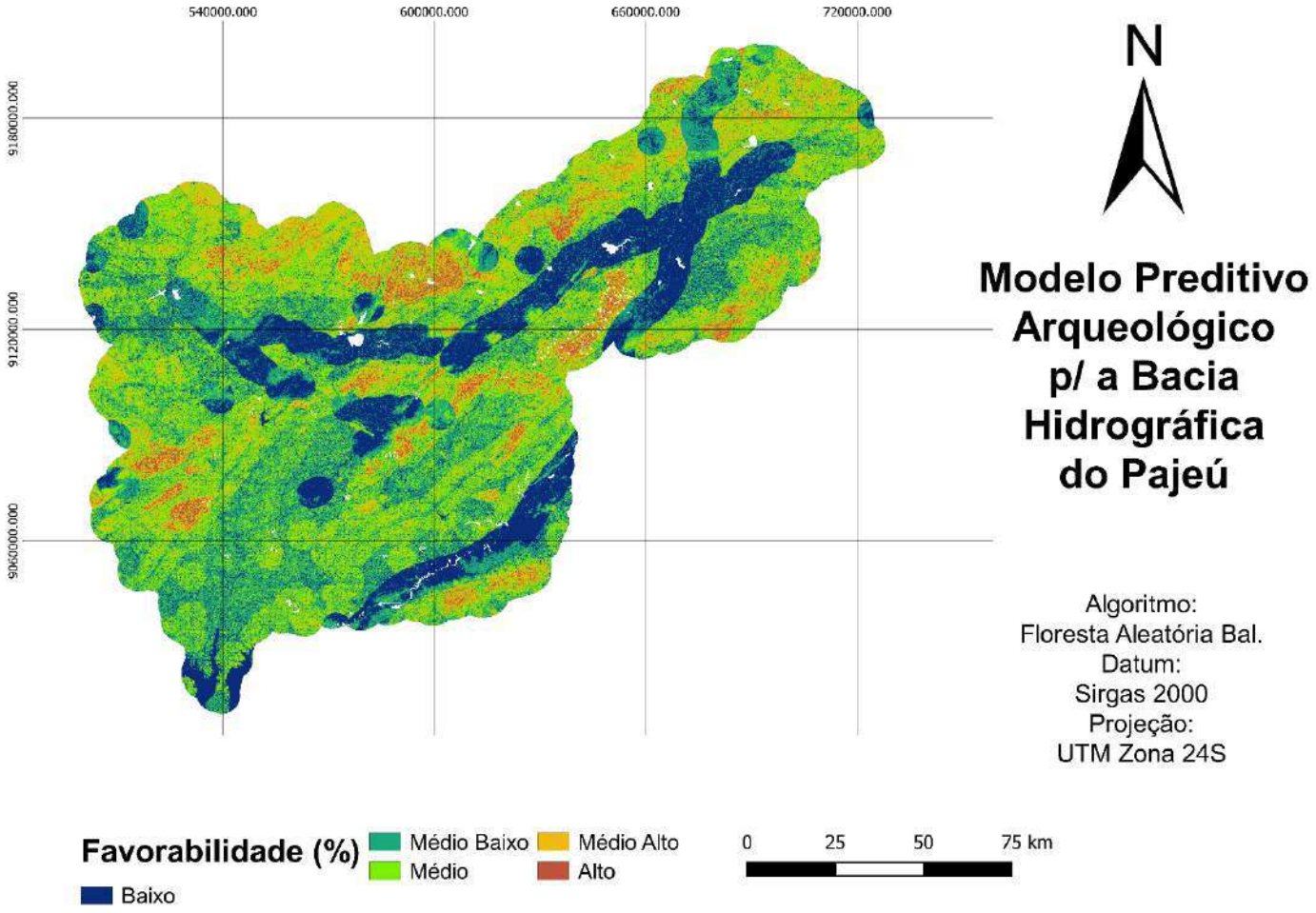
A execução dos modelos de *bagging* balanceado e FAB também foi aqui executada, usando agora os resultados com os dados pós campo e os mapas podem ser vistos nas páginas a seguir (**Figuras 53 e 54**).

Figura 53: MPA_Pajeu_1.0 (*Bagging* balanceado)



Fonte: Elaborado pelo próprio autor (2022)

Figura 54: MPA_Pajeu_1.0 (Floresta Aleatória Balanceada)



Fonte: Elaborado pelo próprio autor (2022)

Os modelos de *bagging* balanceado e FAB continuaram apresentando alta quantidade áreas amarelas (médias), o que pode ser explicado não só pela falta de informação, mas também pelo uso da estratégia de subamostragem. Nesses mapas acima, diferente dos apresentados no capítulo anterior, as áreas de alta probabilidade estão melhor definidas e claramente a proximidade aos CMC baseados na declividade tem uma influência na construção dos produtos do modelo.

As métricas estatísticas de avaliação para os modelos acima (**Tabela 12**) mostram que a qualidade dos resultados, sobretudo para encontrar novos valores positivos baseados nos conjuntos de teste, é positiva. O problema aqui segue sendo a grande quantidade de valores intermediários dentro das probabilidades.

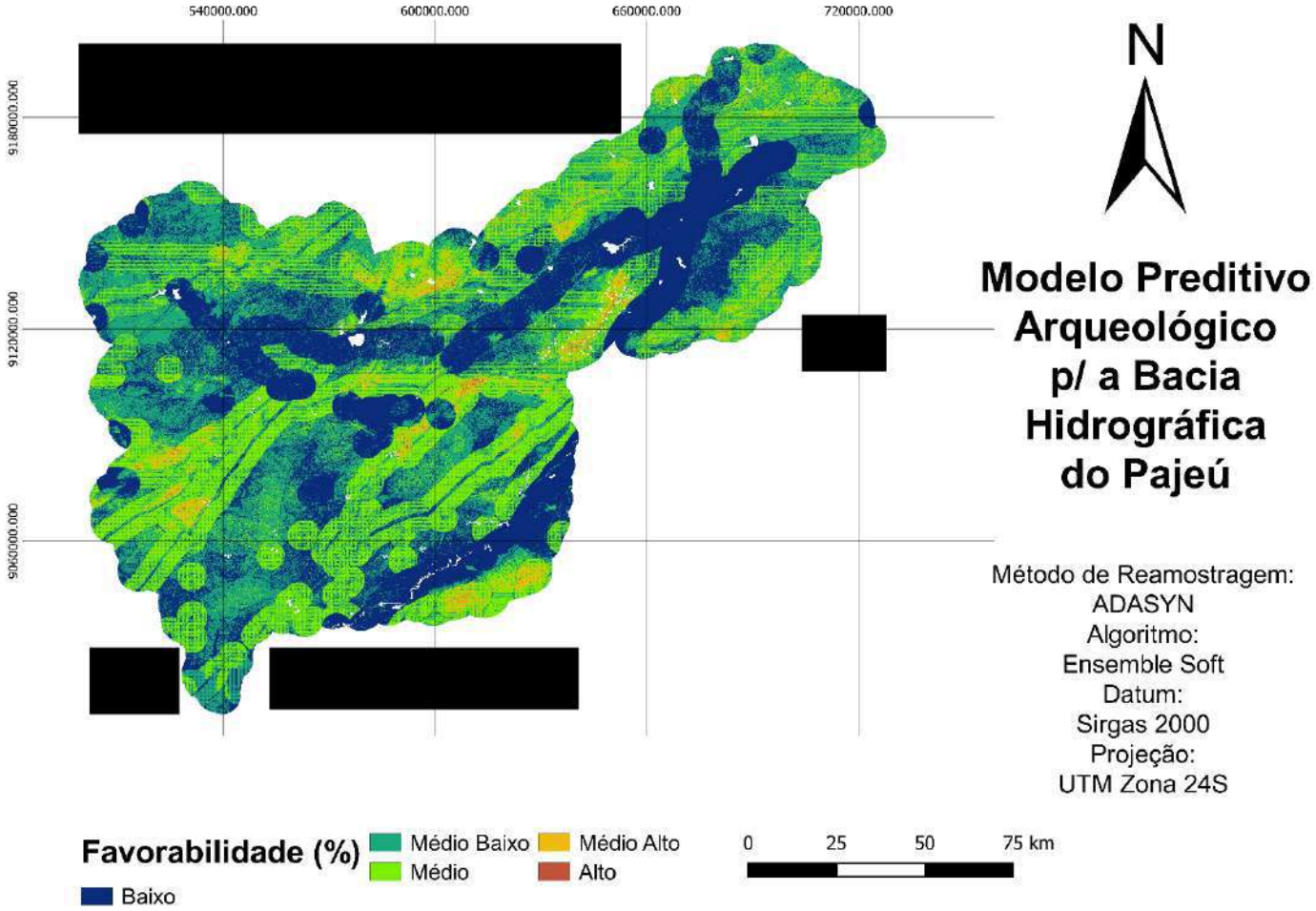
Tabela 12: Métricas estatísticas de avaliação para os modelos finais de *bagging* balanceado e FAB.

	***Bagging* Bal. (BAGG)**	**Floresta Aleatória Bal. (FAB)**
F2 Score	0.827	0.833
Precisão	0.677	0.548
Recall	0.827	0.958

Fonte: Elaborado pelo próprio autor (2022)

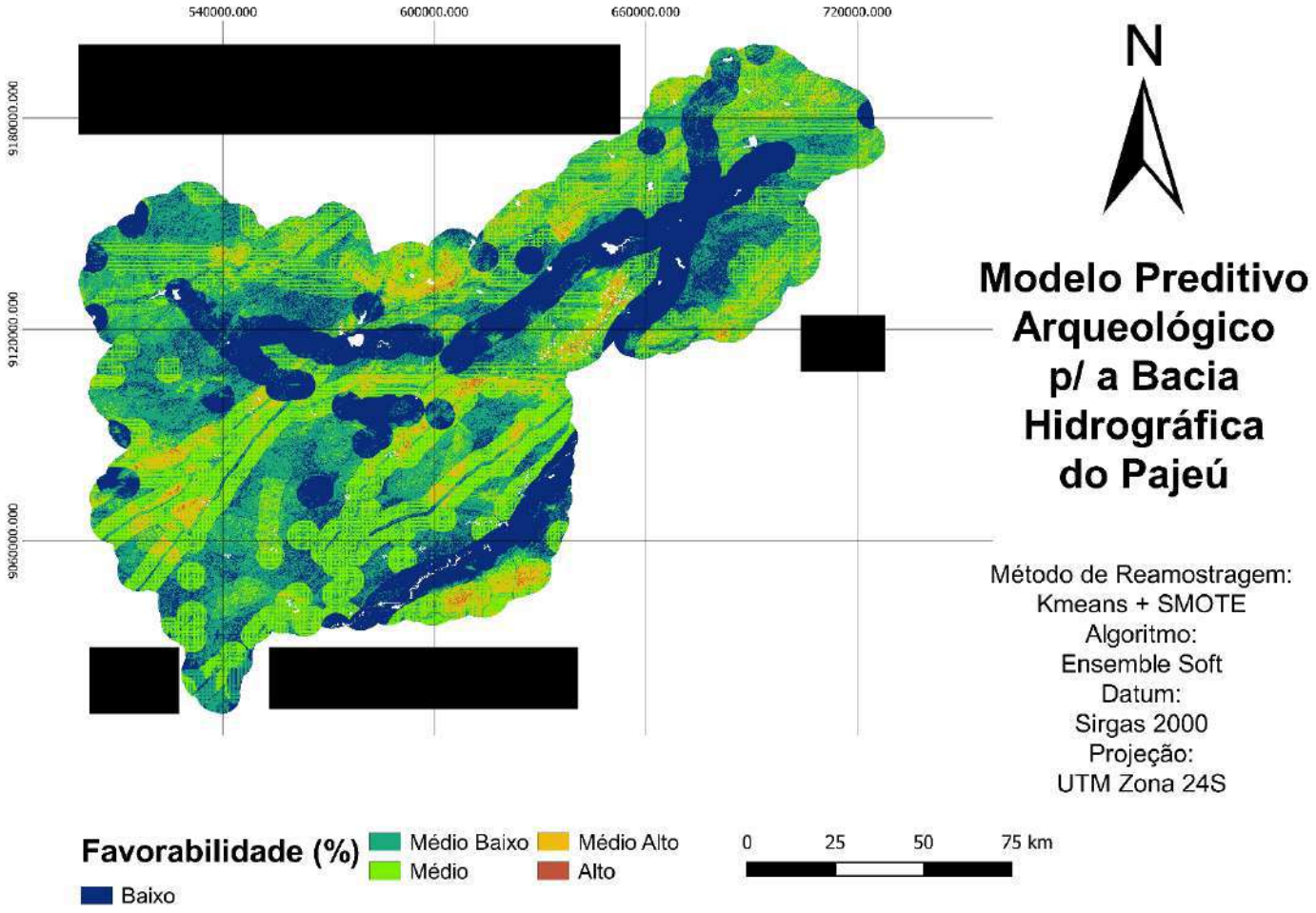
Por fim, os modelos foram somados novamente, de modo a construir o *Ensemble*. Com a junção de todos os modelos chegou-se a um resultado onde, novamente, às áreas de média probabilidade se destacam. Os poucos locais com manchas vermelhas (áreas de alta probabilidade), estão localizados nos mesmos ambientes já relatados, como mostrado nas figuras abaixo (**Figuras 55, 56 e 57**)

Figura 55: MPA_Pajeu_1.0 com o método de reamostragem ADASYN (Ensemble)



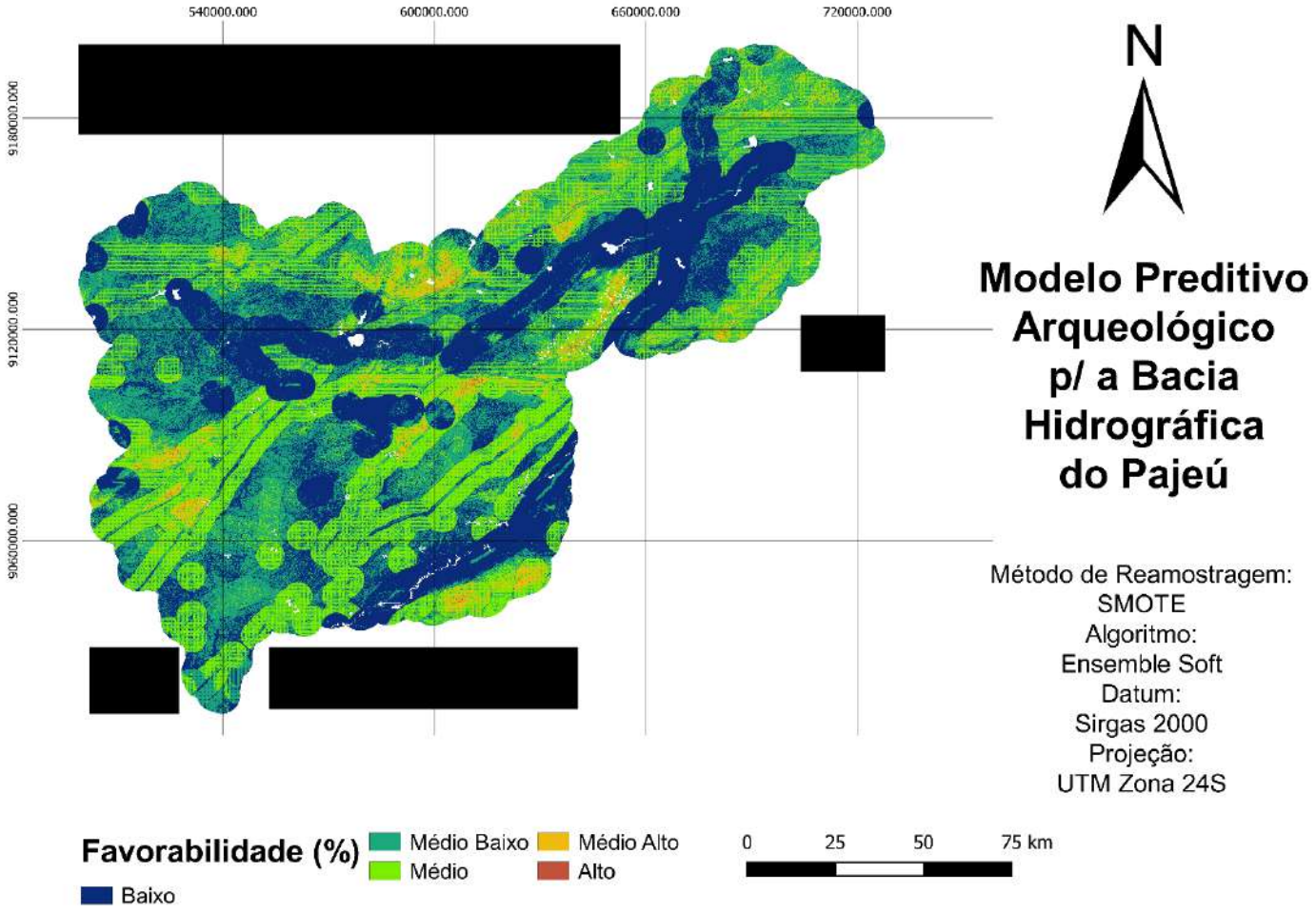
Fonte: Elaborado pelo próprio autor (2022).

Figura 56: MPA_Pajeu_1.0 com o método de reamostragem Kmeans + SMOTE (Ensemble)



Fonte: Elaborado pelo próprio autor (2022)

Figura 57: MPA_Pajeu_1.0 com o método de reamostragem SMOTE (Ensemble)



Fonte: Elaborado pelo próprio autor (2022)

Com relação a avaliação estatística (**Tabela 13**), o modelo segue apresentando valores altos de recall, mas ainda assim às áreas amarelas seguem dominando o mapa. O que fica claro, observando todos os modelos executados, é que, visualmente, o mapa criado através do algoritmo de RNA, teve a melhor configuração, já que os locais de dúvidas são os menores e ainda assim ele consegue ter boas métricas avaliativas. Ainda assim é nítida a falta de informações existentes para o modelo atual. Como resumo dos valores apresentados, é possível observar na **Tabela 14** todas as matrizes de confusão geradas.

Tabela 13: Métricas estatísticas de avaliação para o modelo ensemble.

Ensemble			
	Adasyn	Kmeans Smote	Smote
F2 Score	0.890	0.935	0.924
Precisão	0.955	0.852	0.957
Recall	0.875	0.958	0.917

Fonte: Elaborado pelo próprio autor (2022)

Tabela 14: matriz de confusão dos modelos finais

Smote

	Ausência	Presença
Ausência	164	26
Presença	1	23

Adasyn - Regressão Logísitica

	Ausência	Presença
Ausência	162	28
Presença	2	22

Kmeans Smote

	Ausência	Presença
Ausência	164	26
Presença	2	22

Smote

	Ausência	Presença
Ausência	183	7
Presença	3	21

Adasyn -RNA (Perceptron Multicamadas)

	Ausência	Presença
Ausência	187	3
Presença	3	21

Kmeans Smote

	Ausência	Presença
Ausência	183	7
Presença	3	21

Smote

	Ausência	Presença
Ausência	189	1
Presença	3	21

Adasyn - Floresta Aleatória

	Ausência	Presença
Ausência	188	2
Presença	3	21

Kmeans Smote

	Ausência	Presença
Ausência	190	0
Presença	9	15

Smote

	Ausência	Presença
Ausência	189	1
Presença	2	22

Adasyn - Ensemble

	Ausência	Presença
Ausência	189	1
Presença	3	21

Kmeans Smote

	Ausência	Presença
Ausência	186	4
Presença	1	23

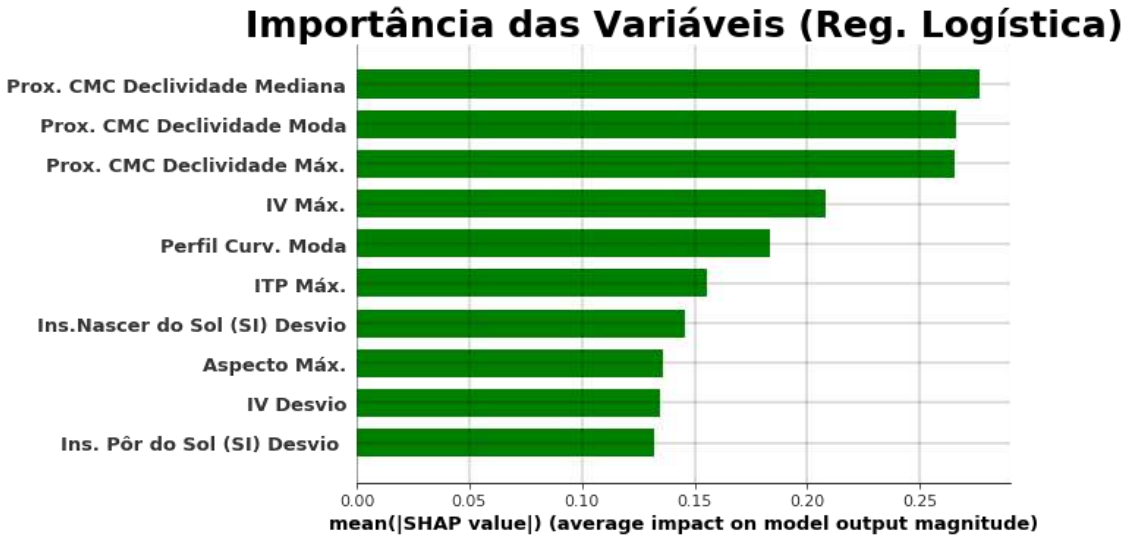
	Bagging Bal.			Floresta Aleatória Bal. (FAB)	
	Ausência	Presença		Ausência	Presença
Ausência	180	10		171	19
Presença	3	21		1	23

Fonte: Elaborado pelo próprio autor (2022)

6.2 A importância das variáveis no MPA_Pajeu_1.0

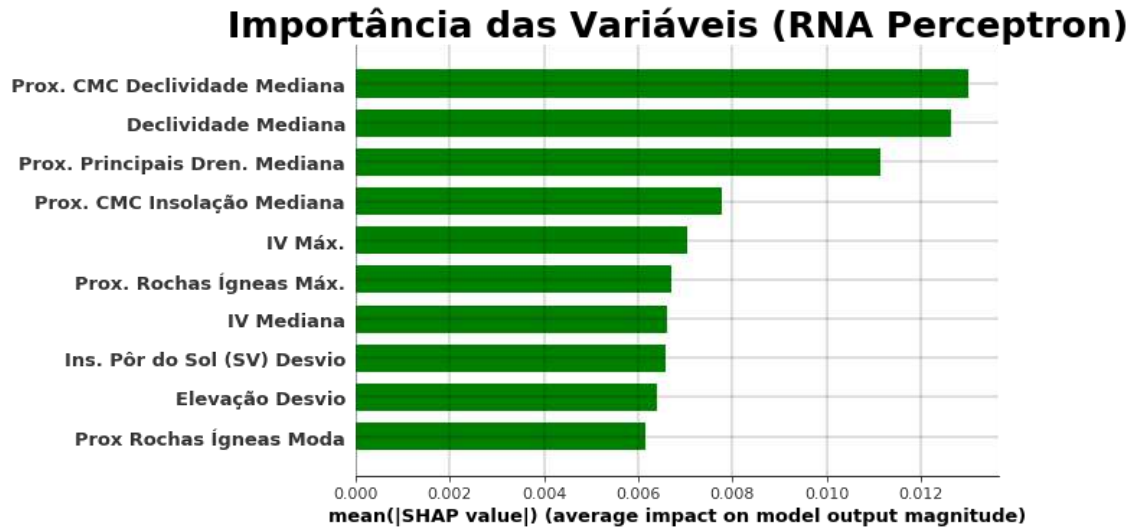
Assim como foi feito para o MPA_Pajeu_0.1, aqui também foram criados gráficos de importância de variáveis usando os valores *SHAP*. Com eles é possível interpretar de uma melhor forma como as variáveis participaram nos modelos finais e se os elementos adicionados pós-campo tiveram algum efeito diferente.

Figura 58: Gráfico de valores *SHAP* de importância das variáveis (Regressão Logística)



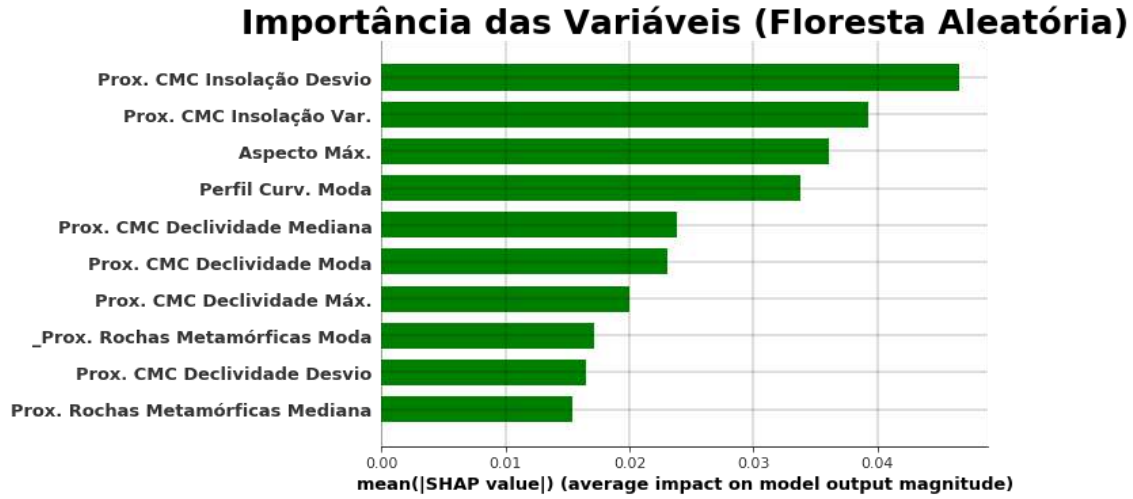
Fonte: Elaborado pelo próprio autor (2022)

Figura 59: Gráfico de valores *SHAP* de importância das variáveis (RNA Perceptron)



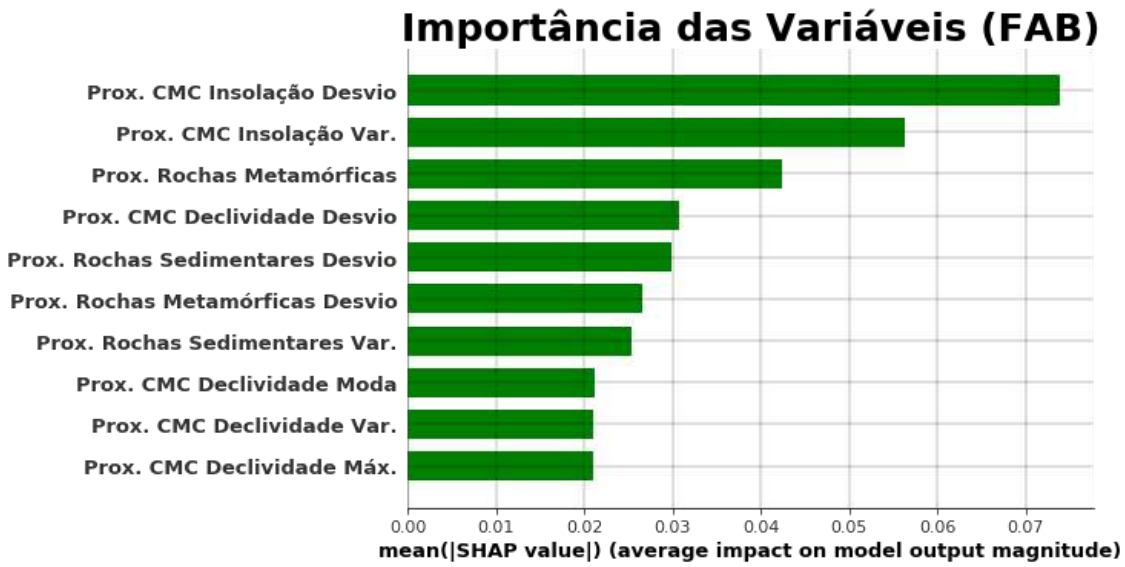
Fonte: Elaborado pelo próprio autor (2022)

Figura 60: Gráfico de valores *SHAP* de importância das variáveis (Floresta Aleatória)



Fonte: Elaborado pelo próprio autor (2022)

Figura 61: Gráfico de valores *SHAP* de importância das variáveis (FAB)



Fonte: Elaborado pelo próprio autor (2022)

Olhando os gráficos acima (**Figuras 58 a 61)** é possível perceber que de fato as variáveis derivadas da acessibilidade tiveram uma importância crucial nos modelos, sobretudo no que se refere a proximidade aos CMC que usaram a declividade como superfície de custo. É possível ver que ela aparece em todos os diferentes algoritmos.

Com relação ao Aspecto máximo, que já vinha sendo destaque nos modelos de Floresta Aleatória e Regressão logística do modelo inicial, o mesmo continua aparecendo aqui também de forma destacada. Além deste, ao olharmos o modelo de regressão logística, é possível ver o IV máximo também em destaque, assim como no inicial. O perfil de Curvatura também reaparece nos modelos de Floresta Aleatória.

O grande destaque, porém, é o modelo de FAB. Ao olharmos para as principais variáveis que contribuem no produto final do modelo, vemos que dois fatores foram essenciais, uma foram as variáveis derivadas da acessibilidade, sejam elas com insolação ou declividade como superfície de custos, mas também a proximidade para áreas de rochas metamórficas e sedimentares, o que coloca a questão do suporte rochoso em evidência.

Analisando de forma geral os resultados obtidos nós podemos perceber que há algumas variáveis essenciais para construção desses modelos finais, sobretudo as referentes à acessibilidade, que de certa forma, só de analisar os mapas, indicam que os sítios, em sua maioria, não estão em locais com grande densidade de rotas, mas sim na periferia destas, em locais topograficamente elevados em relação ao entorno, o que pode explicar também a presença do ITP ou da própria elevação aparecendo como variáveis significativas em alguns modelos.

O fato destes não estarem associados a estes caminhos principais, não anula a ideia de usar os grafismos como um sistema comunicador e a associação a variáveis como o aspecto, podem indicar que a localização destes pode estar associada a locais de práticas rituais pelos grupos, ou até mesmo estarem relacionados a conservação dos sítios já conhecidos e que foram usados como base para a construção do modelo.

O modelo de RNA, que não havia sido usado na primeira fase, aqui fornece resultado importantes também, mostrando que a insolação e a proximidade para os cursos d'água também contribuem para achar os sítios de grafismos rupestres. Talvez por isso estes modelos tiveram um melhor resultado em observar áreas de possíveis sítios arqueológicos nas áreas mais rebaixadas do Pajeú.

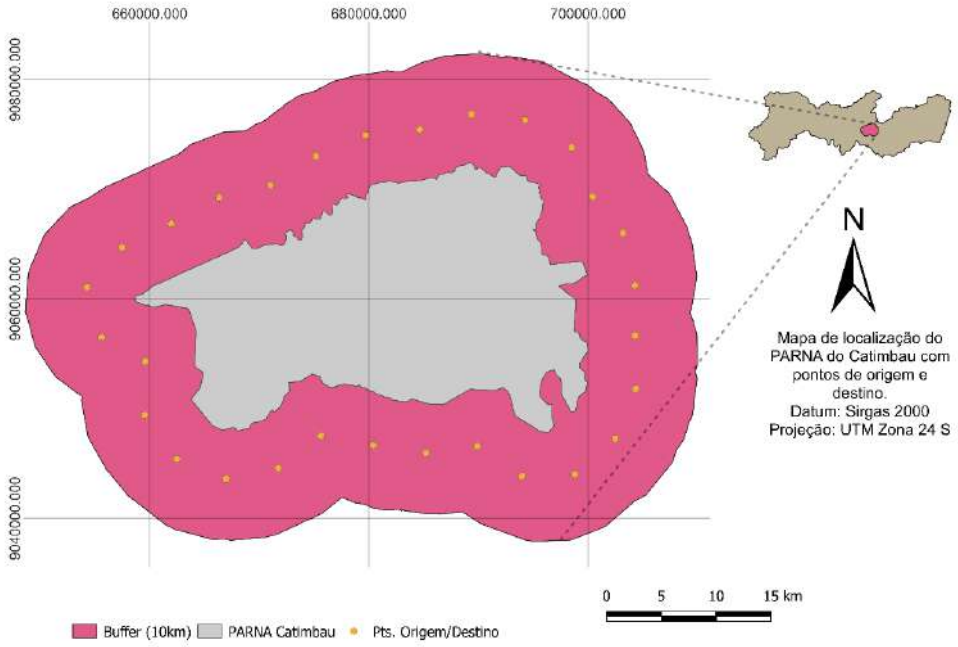
7. PONDO OS MODELOS À PROVA

Além de toda avaliação estatística apresentada nos capítulos anteriores é preciso que os modelos sejam testados na prática , seja indo a campo, ou também aplicando os produtos gerados em áreas já intensamente estudadas, as quais podem ser testadas à luz dos dados treinados na Bacia do Pajeú.

Para esse primeiro caso os modelos treinados no Pajeú puderam ser aplicados em um outro local do semiárido pernambucano através da Transferência de Aprendizado (Referências), que busca aplicar um modelo treinado para um determinado problema em outra perspectiva ou para resolver outro tipo de problemática. No caso desta tese escolheu-se a região do PARNA do Catimbau, local de alta densidade de estudos e sítios arqueológicos conhecidos, para ser usada como um teste prático.

O PARNA do Catimbau abarca uma área de 62.294 Ha (ICMBio, 2022) e conta com 15 sítios cadastrados no SICG/IPHAN. Para o cálculo do modelo preditivo da área, utilizou-se um *buffer* de 10km de borda além dos limites do Parque, onde foram inseridos 29 pontos de partida, com distâncias de 5km entre cada um deles, como mostrados no mapa abaixo (**Figura 62**), a fim de calcular os CMC.

Figura 62: Localização do PARNA do Catimbau e locação dos pontos de origem/destino.



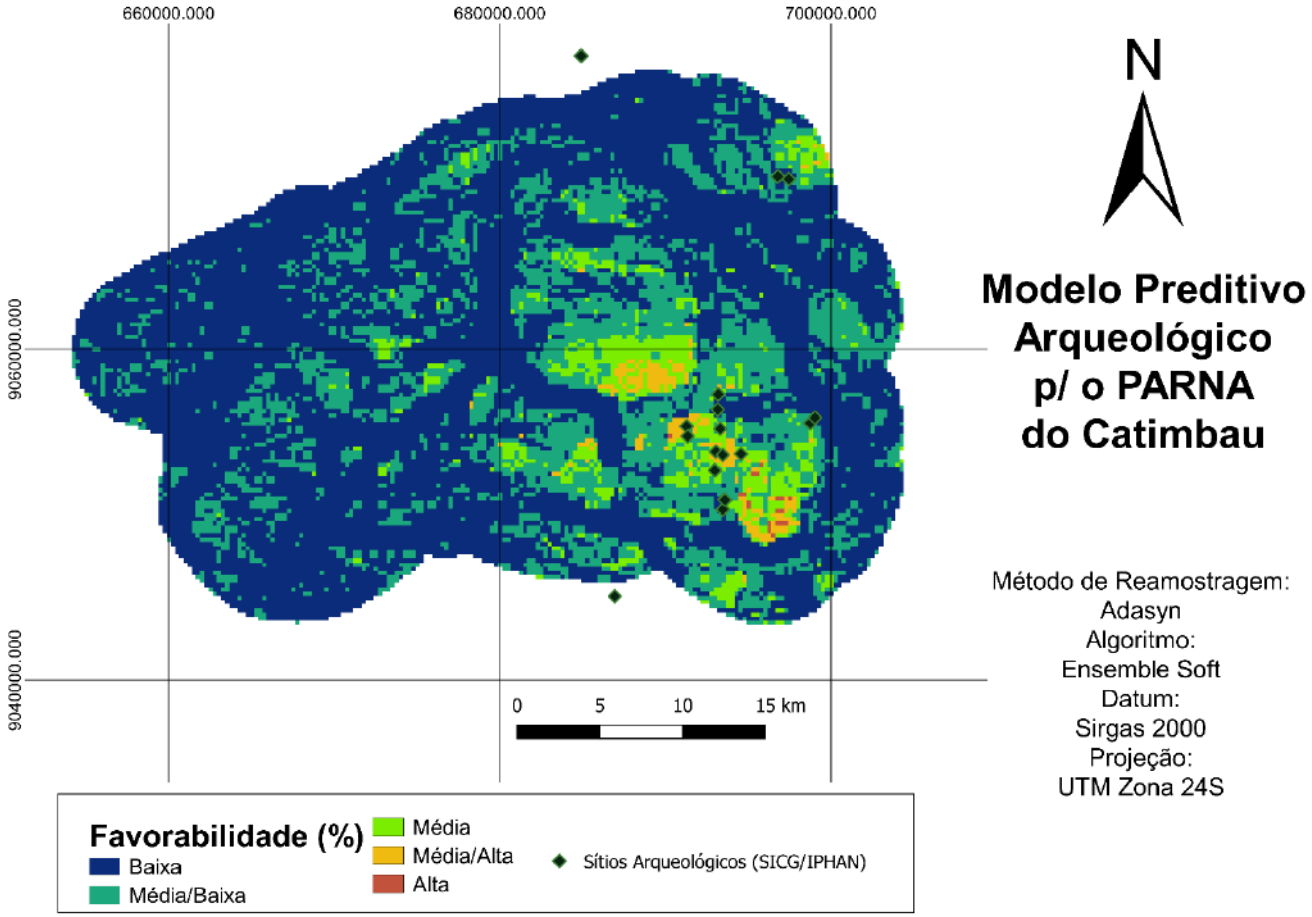
Fonte: Elaborado pelo próprio autor (2022)

7.1 Calculando as estatísticas do modelo

Assim como feito para a área do Pajeú em que ainda não se tinham informações acerca de presença ou ausência de sítios, aqui se usou toda área de Parque e seu entorno imediato dentro de um contorno de 5 km de borda, para se calcular as estatísticas referentes as mesmas variáveis usadas para o Pajeú, considerando toda área analisada como uma grande mancha cinza, portanto não levando em consideração nenhuma informação sobre os sítios da região.

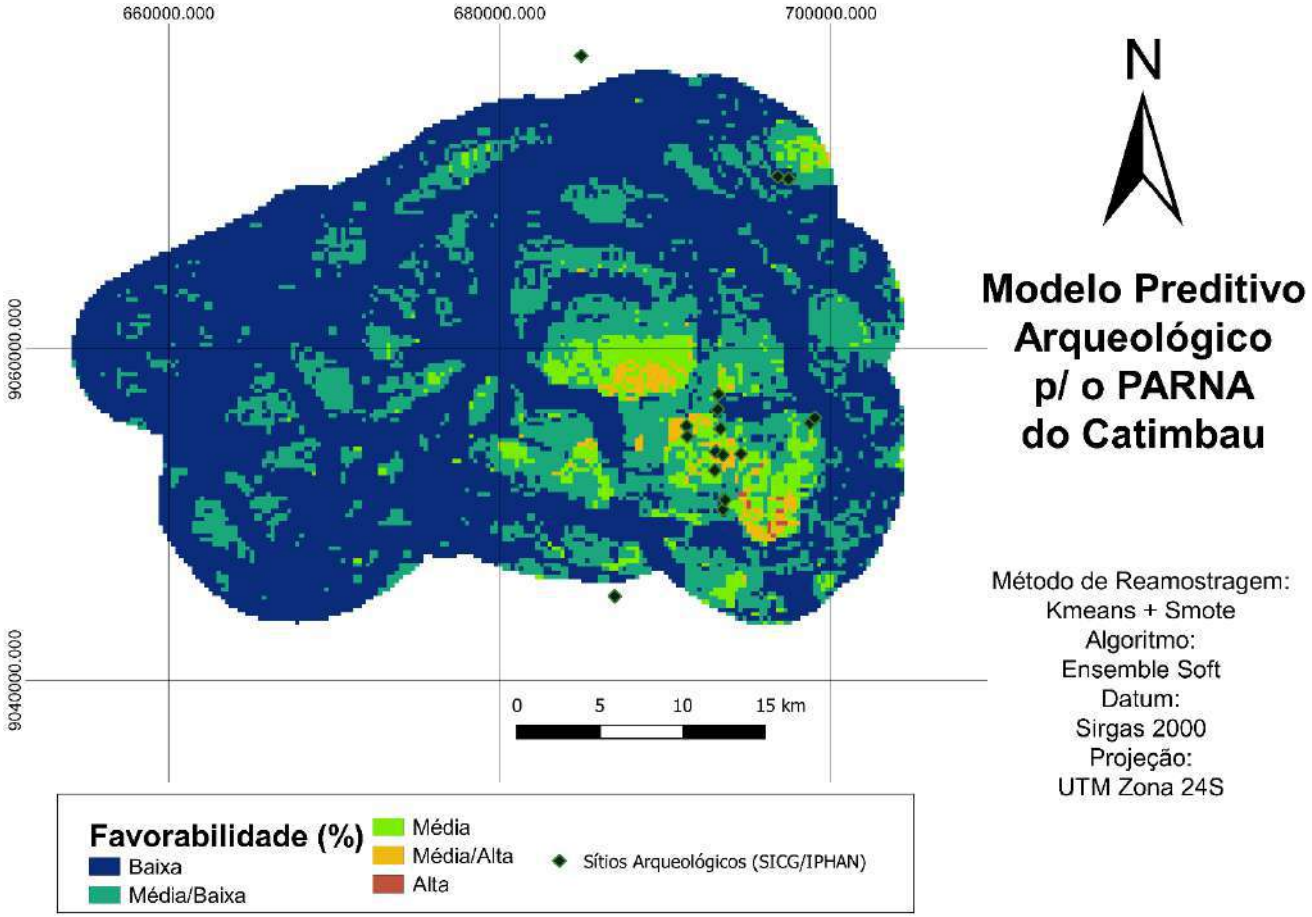
Ao final do cálculo, assim como já havia sido feito para os modelos anteriores, todos os dados foram importados para o *python* para enfim serem testados, tendo como base o modelo treinado na região da Bacia do Pajeú. Os resultados obtidos podem ser vistos nos mapas abaixo (**Figuras 63, 64 e 65**), em que focamos nos modelos ensemble para as três categorias de reamostragem.

Figura 63: MPA PARNA do Catimbau com o método de reamostragem Adasyn (Ensemble)



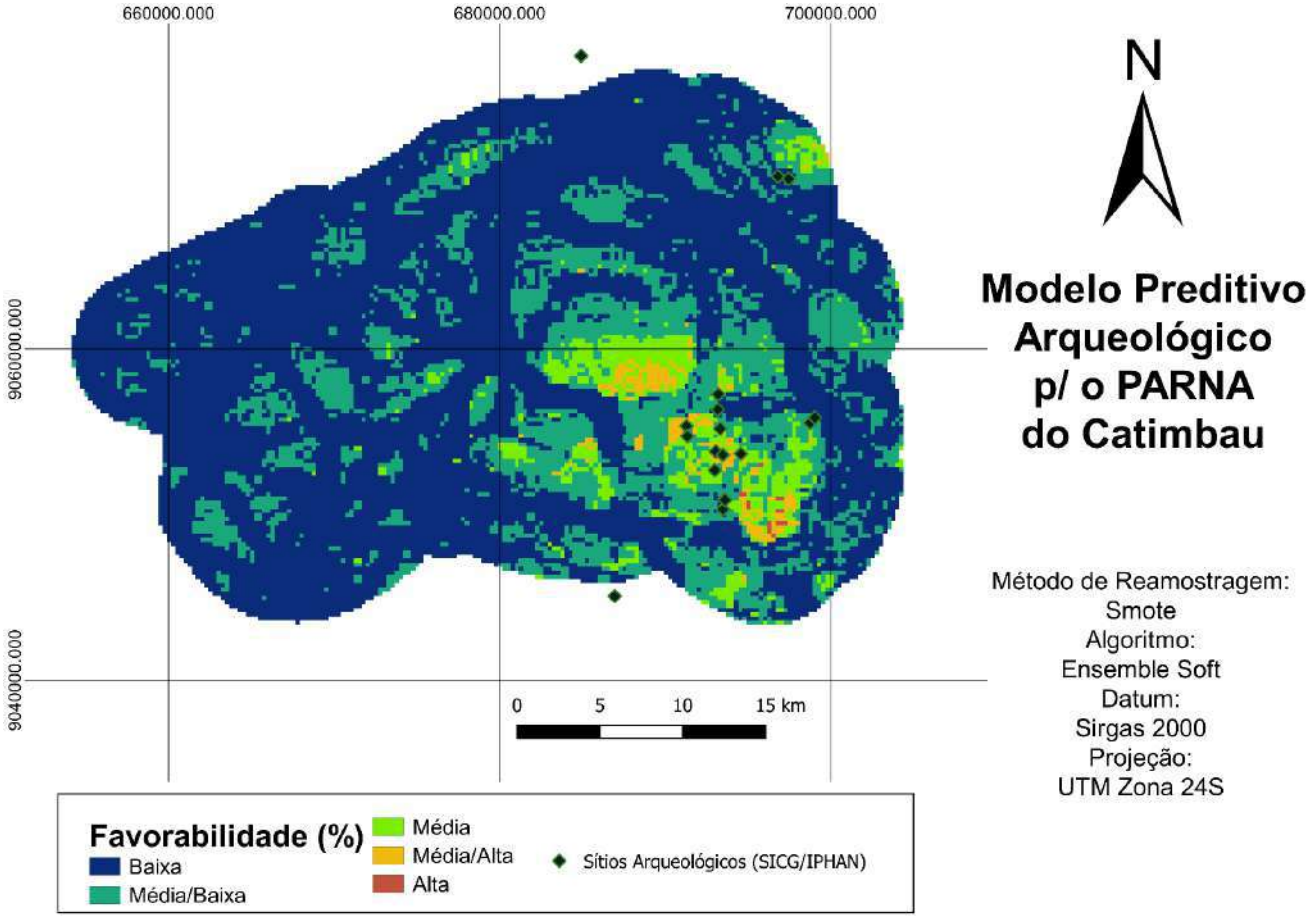
Fonte: Elaborado pelo próprio autor (2022)

Figura 64: MPA PARNA do Catimbau com o método de reamostragem Kmeans + Smote (Ensemble)



Fonte: Elaborado pelo próprio autor (2022)

Figura 65: MPA PARNA do Catimbau com o método de reamostragem Smote (Ensemble)



Fonte: Elaborado pelo próprio autor (2022)

7.2 Discutindo os resultados

Como é possível ver nos mapas acima todos os modelos, não importando o método de reamostragem utilizado, tiveram configurações muito similares com relação a distribuição das áreas de favorabilidade. Ao contrário do que fizemos para os modelos mais completos, no caso do Catimbau, optou-se por calcular todos os algoritmos e assim gerar o modelo ensemble no final para então poder analisar os resultados.

Fica claro que as áreas de alta e média alta favorabilidade estão bem relacionadas com os sítios da base do SICG/IPHAN, atestando uma boa performance do produto treinado no Pajeú e indicando que o mesmo pode ser transferido para outras regiões, ao menos do semiárido nordestino, as quais tem uma configuração ambiental similar.

.

7.3 A segunda visita a campo

No fim do ano de 2022, já com a pandemia de Covid-19 em declínio, foi possível realizar uma nova visita a campo, nesta utilizando uma metodologia mais sistemática de uso dos produtos gerados para busca de novos sítios arqueológicos na Região do Pajeú.

Foram visitadas seis áreas diferentes, três no norte da Bacia, entre os municípios de Flores, Afogados da Ingazeira e Iguaraci, indo a locais de baixa, média e alta probabilidade e tendo como base também sítios arqueológicos relatados por moradores locais que ainda não estavam na base de dados utilizada.

No município de Afogados, mais precisamente na Serra do Pereiro (**Figura 66**), foi recebido uma informação de presença de um sítio arqueológico com grafismos rupestres no topo de uma elevação. Realizamos, pois, uma prospecção neste local, que foi previsto com uma área de alta probabilidade, porém, devido às dificuldades impostas pelo terreno e a vegetação muito densa por conta das chuvas, foi impossível chegar até os vestígios em si.

Figura 66: Serra do Pereiro

Fonte: Foto tirada pelo próprio autor (2022)

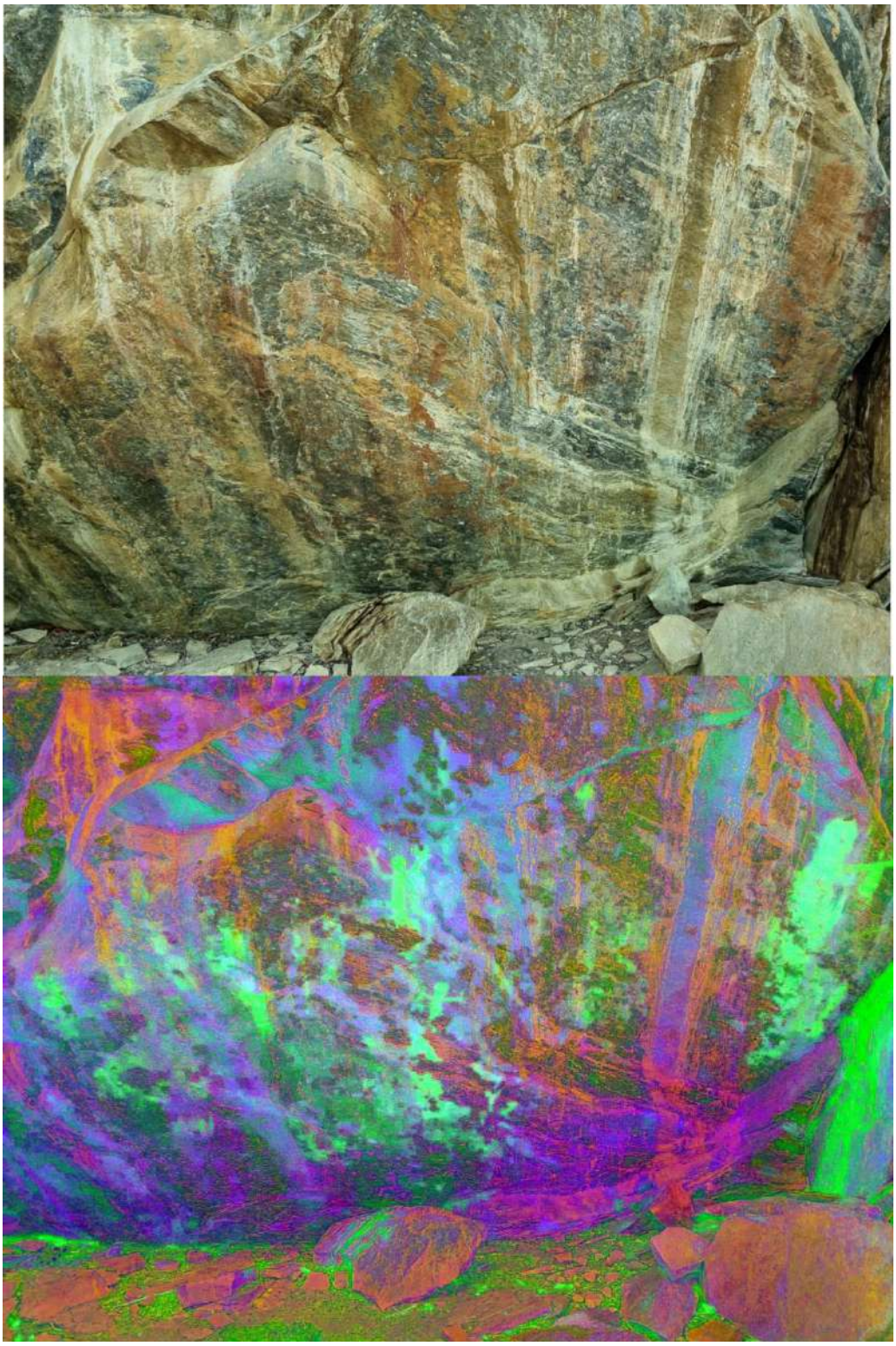
A segunda área de prospecção visitada foi no município de Flores, às margens do Rio Pajeú. Aproveitando-se da presença conhecida de um sítio numa área de média probabilidade, definiu-se um polígono de prospecção ao longo das margens do Rio Pajeú. Foram visitados alguns afloramentos dentro dessa área, mas nenhum outro registro, além do sítio Pedra do Riacho da Velha (**Figura 67**), foi encontrado.

Figura 67: Sítio Pedra do Riacho da Velha

Fonte: Foto tirada pelo próprio autor (2022)

A última área visitada na região fica entre os municípios de Iguaraci e Afogados da Ingazeira, próximo ao sítio Serrote Verde (**Figura 68**), local de baixa probabilidade. Foram observados outros afloramentos além do local onde está o sítio, porém não foi encontrado nenhum outro grafismo na área.

Figura 68: (em cima) Imagem geral do suporte com pinturas no sítio Serrote Verde. (abaixo) Imagem tratada no aplicativo Rock Art Enhancer, onde é possível ver claramente os antropomorfos dispostos no painel.



Fonte: Foto tirada pelo próprio autor (2022)

As outras três áreas visitadas durante as atividades realizadas em 2022 foram novamente entre os municípios de Carnaubeira da Penha e Mirandiba, anteriormente visitados. Desta vez, porém, foram encontrados mais dois sítios que ainda não estão inclusos na base usada para treinar os modelos criados.

Primeiramente foi visitada uma área ao longo do leito do Riacho do Brejo do Gama, um dos afluentes do Rio Pajeú, no município de Mirandiba (**Figura 69**). Esta área é calculada pelo modelo como um local de média ou baixa probabilidade predominantemente. Nela foi encontrado um sítio arqueológico com gravuras e também foi registrado um outro sítio, também de gravuras, mas que já se encontra na base de dados do modelo.

Figura 69: à esquerda: Sítio Lajedo dos Porfírios, ainda não cadastrado; à direita sítio Pedra Gravada, já presente na base de dados

Fonte: Foto tirada pelo próprio autor (2022)

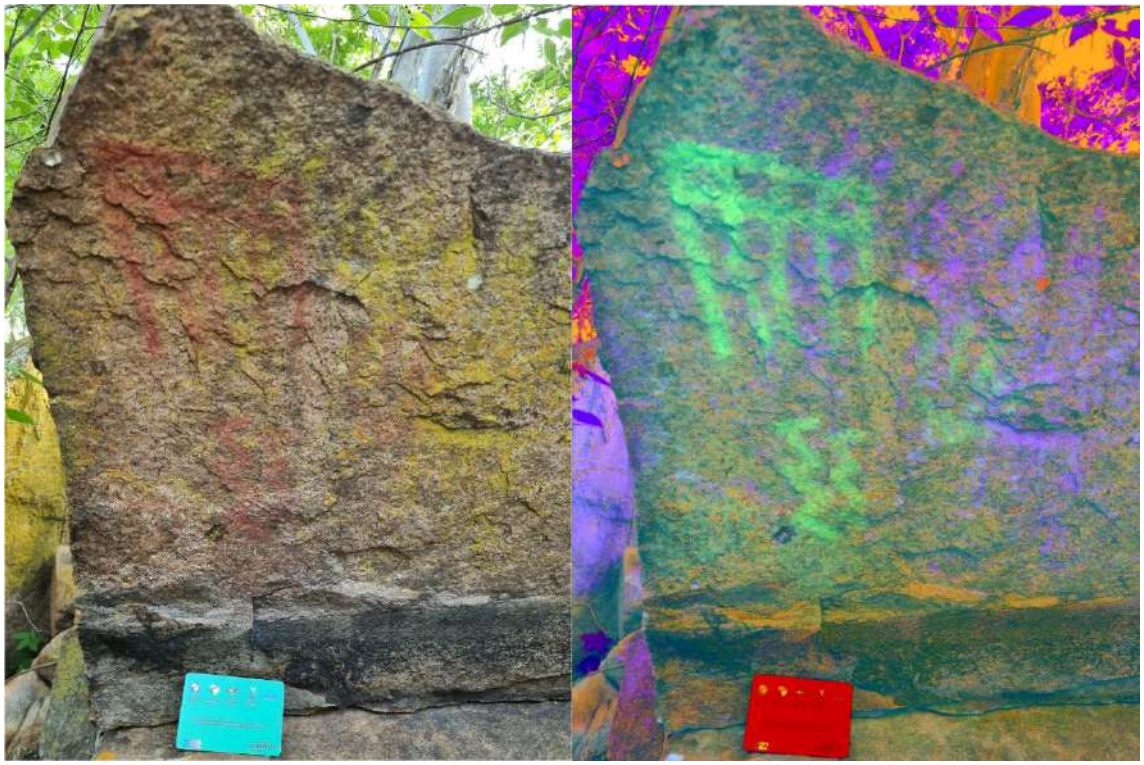
A segunda área visitada foi a Serra da Tiúba, em Carnaubeira da Penha, local de alta probabilidade segundo o modelo, mas onde não foram encontrados nenhum sítio arqueológico. Na área foi possível observar diversas áreas de abrigos, bem similares a outros locais com presenças de registro rupestre, como mostrado na figura abaixo (**Figura 70**).

Figura 70: Abrigo na Serra da Tiúba.

Fonte: Foto tirada pelo próprio autor (2022)

Por fim, o último local visitado foi a Serra das Flores, local também de alta probabilidade, onde foi encontrado um sítio de pinturas que ainda não estava inserido na base dados do modelo (**Figura 71**). O sítio conta com diversas pinturas espalhadas ao longo de diversos afloramentos rochosos, tendo marcas de mão, antropomorfos e formas geométricas.

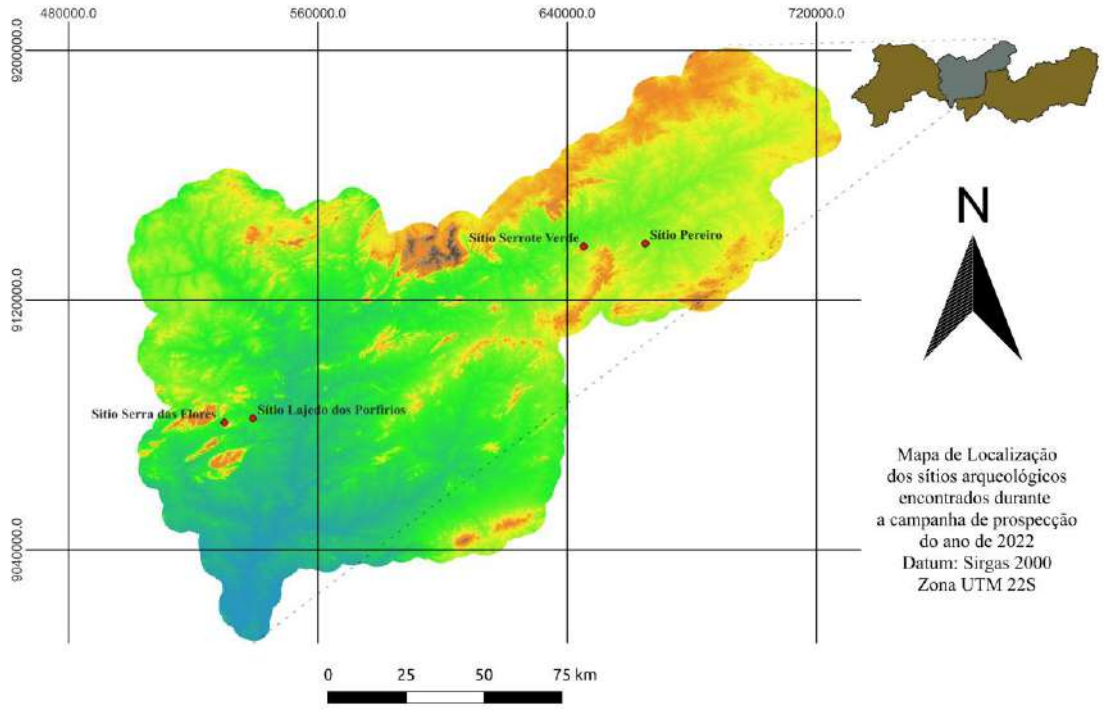
Figura 71: à esquerda: imagem de um dos grafismos encontrados no sítio da Serra das Flores. À direita: foto editada no software Rock Art Enhancer.



Fonte: Foto tirada pelo próprio autor (2022)

É importante informar que ao fim da visita, além da identificação de novos sítios arqueológicos, foi possível também obter dados de caminhamento com o uso de aparelhos GNSS de navegação, os quais poderão ser usados em futuras melhorias no modelo preditivo. A localização de cada um dos sítios pode ser conferida no mapa abaixo (**Figura 72**).

Figura 72: Mapa de localização dos sítios encontrados durante a prospecção realizada no ano de 2022.



Elaborado pelo próprio autor (2023)

.7.4 Discutindo os resultados do campo

Com os dados apresentados acima fica claro que o modelo conseguiu performar bem em prever locais com presença de grafismos rupestres, porém nota-se que determinados padrões ainda não são percebidos, como fica claro com o sítio Serrote Verde, que está localizado numa área proeminente em relação a seu entorno, mas com cotas altimétricas bem reduzidas a ponto das imagens SRTM puderem denotar algo.

Outro ponto interessante deste campo, é que ficou comprovado que mesmo locais que tenham elementos físicos bem robustos e formações rochosas propicias para se encontrar grafismos rupestres, tanto de acordo com o modelo, quanto com o que já se conhece de pesquisas, não são garantia de serem locais com presença de grafismos rupestre, como é o caso

da Serra da Tiúba, que nas duas visitas feitas, tanto em 2021 como em 2022, nenhum grafismo foi avistado.

Essas questões deixam claro que o MPA tem seu papel importante, mas existem diversos elementos que não podem ser percebidos devido a uma limitação dos próprios dados usados para a construção dos produtos, além disso os fatores intempéricos podem já ter atuado tão fortemente em determinados locais visitados, que os registros que ali havia, foram apagados.

.7.5 Separando as gravuras das pinturas

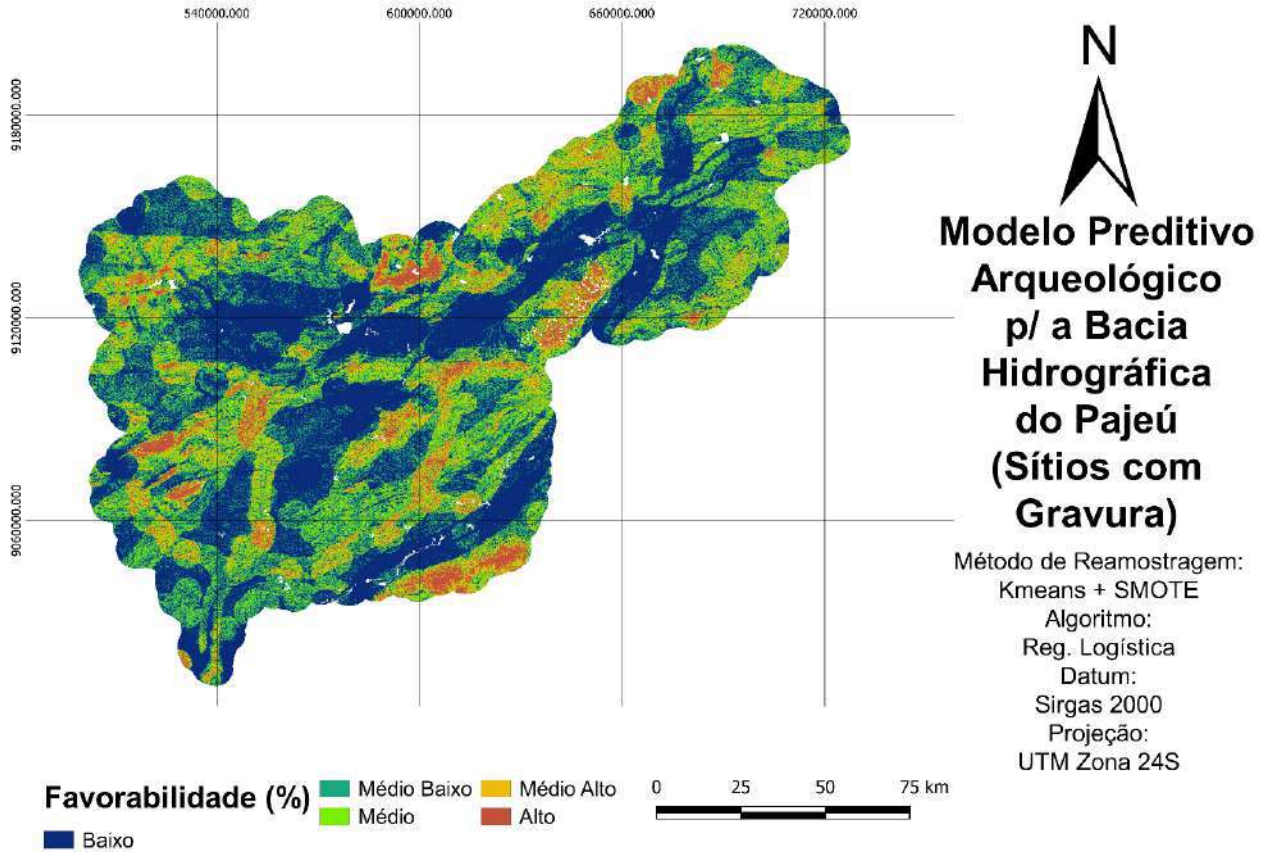
Um dos objetivos deste trabalho previa a criação de um modelo que diferenciasse gravuras de pinturas, já que na literatura é bem comum o relato de uma diferenciação espacial entre as duas técnicas. Inicialmente já é importante frisar que este teste foi feito apenas para observar o comportamento do modelo, porém as estatísticas já indicavam uma baixa confiabilidade dos resultados, com valores de recall e precisão muito baixos, o que muito se deve pela perda de informações, que mesmo ao utilizar os dados em sua completude, ainda se mostravam insuficientes para alguns locais.

Foram gerados diversos mapas, mas para título de comparação aqui só mostraremos os resultados obtidos utilizando a regressão logística e o Kmeans + Smote como método de reamostragem, para cada uma das técnicas (**Figuras 73 e 74**).

Os mapas abaixo seguem uma configuração bem similar aos modelos completos, se olharmos sobretudo as áreas que já vinham aparecendo como locais de alta favorabilidade elas são persistentes aqui também. Não há uma grande diferença entre os modelos de gravura e de pintura, indicando que estas podem ser encontradas nos mesmos ambientes, porém pode-se observar que nas áreas mais rebaixadas, sobretudo no leito do rio Pajeú, é possível encontrar uma alta favorabilidade para gravuras, que não aparece no caso das pinturas.

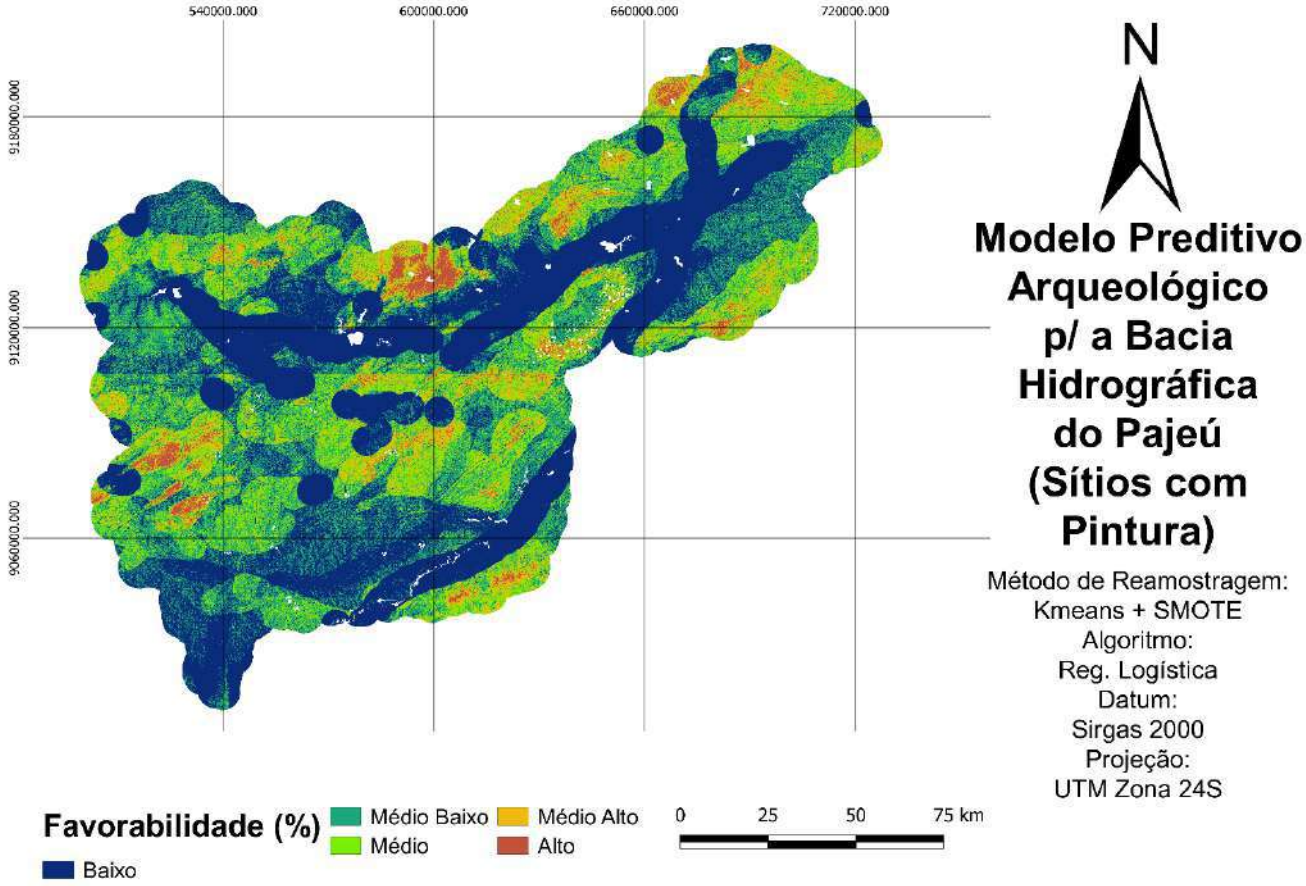
É possível ver também uma grande área de dúvida, o que já era esperado devido a base de dados reduzida, que acaba perdendo informações para a compreensão dos padrões associados. Diante do exposto, mesmo observando configurações de modelos semelhantes, não é possível dizer que gravuras e pinturas serão achados nos mesmos ambiente sempre, só o aumento da base dados vai garantir a construção de modelos mais confiáveis estatística e espacialmente.

Figura 73: MPA treinado exclusivamente com os sítios de gravuras usando o método de reamostragem Kmeans+ SMOTE (Regressão logísitca)



Fonte: Elaborado pelo próprio autor (2022)

Figura 74: MPA treinado exclusivamente com os sítios de gravuras usando o método de reamostragem Kmeans+ SMOTE (Regressão logísitca)



Fonte: Elaborado pelo próprio autor (2022)

.7.6 MPA_Pajeu_1.1: Os modelos pós-defesa

Este subtópico foi escrito após as considerações da banca durante a defesa, quando foi sugerido que fossem feitas algumas modificações nos dados usados para calcular os modelos. A escolha por criar este novo tópico e não substituir aos dados apresentados na versão anterior foi proposital, pois isso só mostra como funciona a construção de um modelo preditivo, a cada nova inserção ou mudanças na base dados, novas versões surgem.

Com relação as alterações desta nova versão, foi sugerido primeiramente que houvesse uma redução das variáveis, pois foram usadas cinco métricas estatísticas nas versões anteriores (mediana, desvio padrão, variância, valor máximo e a moda). Nestes novos modelos, porém, apenas a mediana e o desvio padrão foram calculados.

Além disso, a variável de acessibilidade foi reduzida a apenas ao uso da superfície de custo e não mais do caminho em si. Esta superfície de custo é a base para o cálculo dos caminhos de menor custo, já que é um raster que contém os valores a serem gastos para se mover em cada um dos pixels que compõem a imagem da área a ser calculada. Ao final uma tabela de atributos com apenas 46 variáveis foi usada para gerar os novos modelos.

Como a maioria dos resultados quando comparamos os diferentes métodos de reamostragem foi novamente muito similar, aqui iremos mostrar os mapas utilizando o SMOTE, como técnica para lidar com dados desbalanceados, além disso também serão mostrados os produtos gerados com os algoritmos de FAB e Bagging balanceado.

Figura 75: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (Regressão logística)

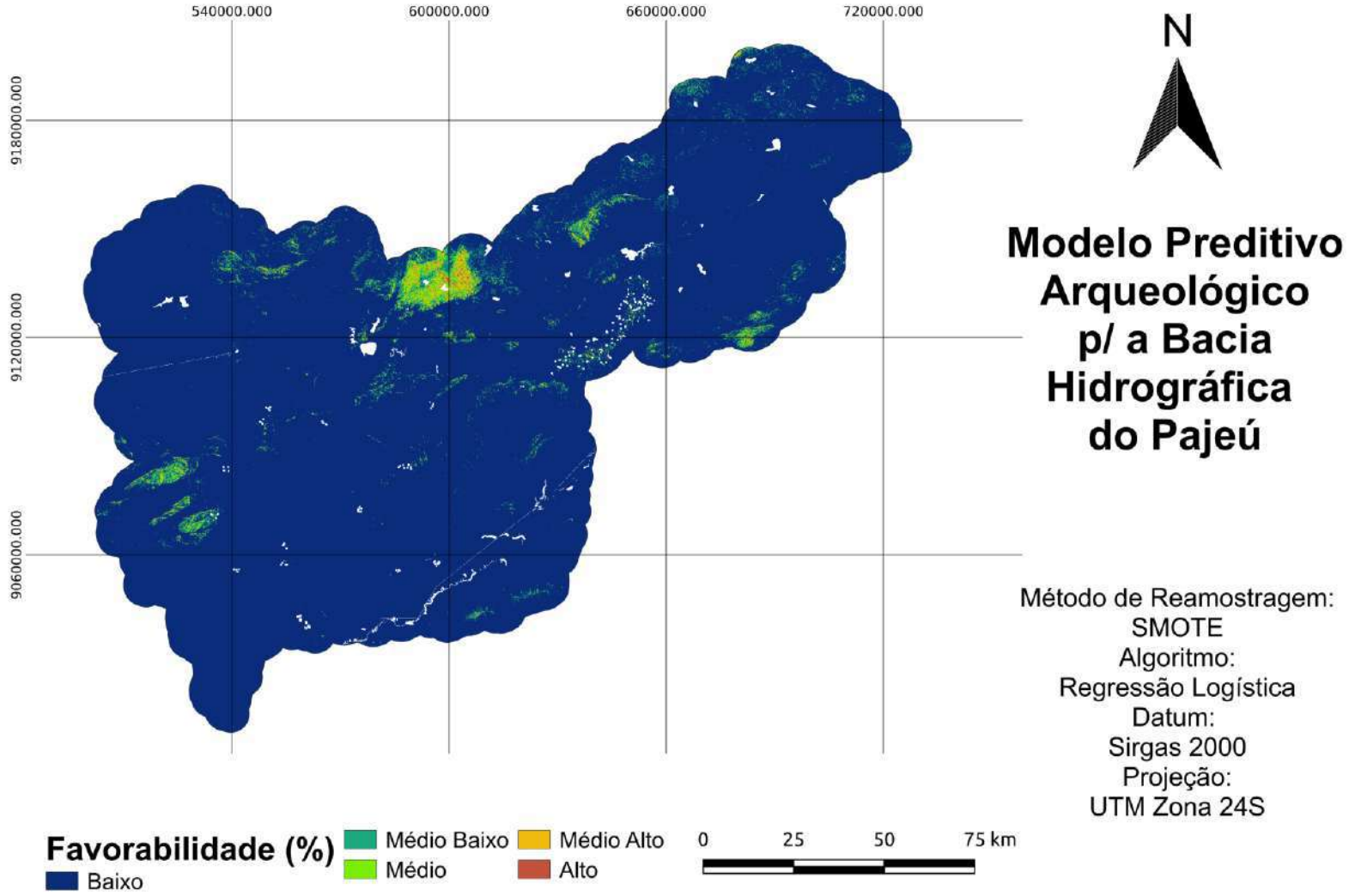


Figura 76: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (RNA – Perceptron MC)

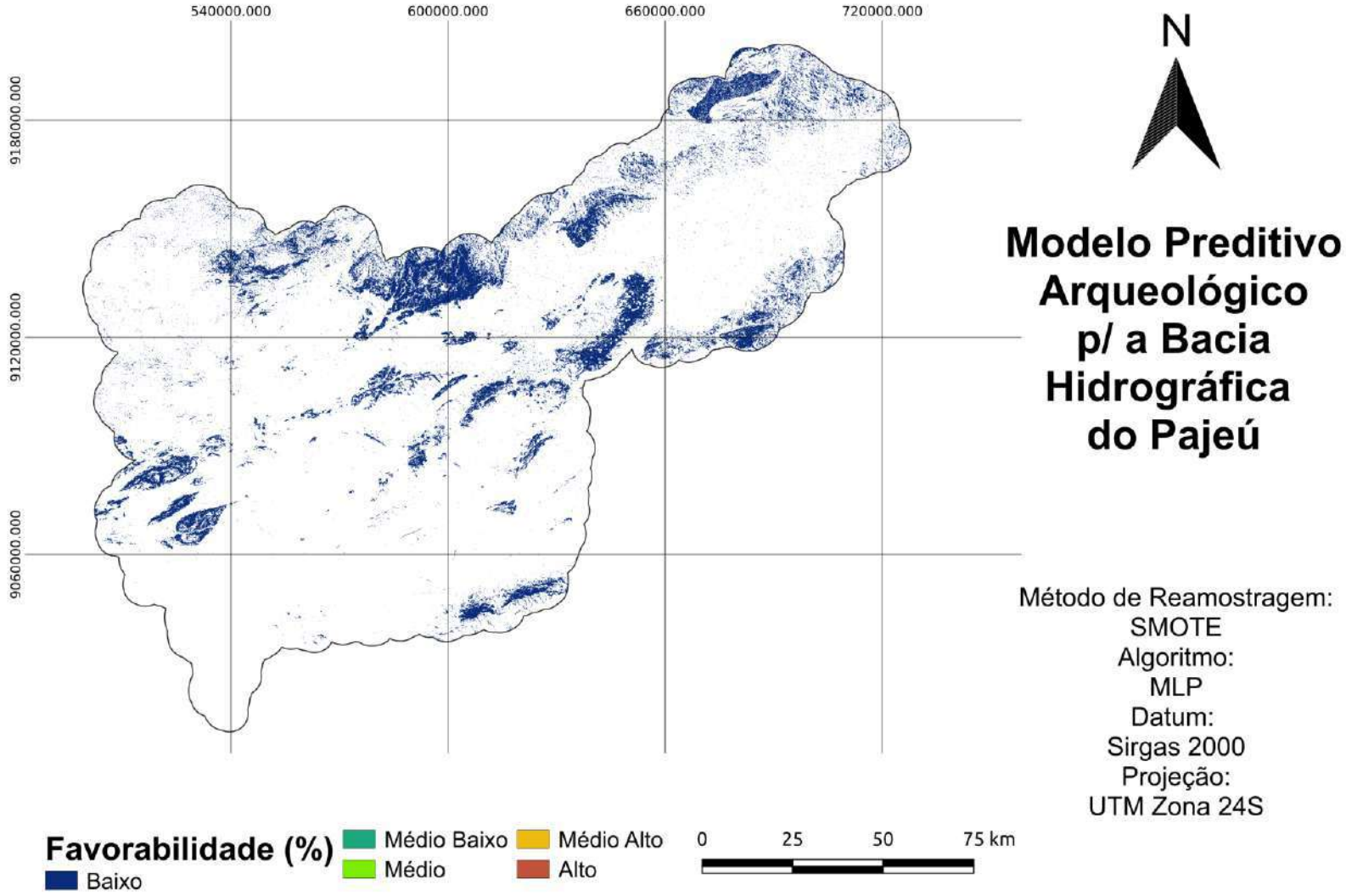


Figura 77: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (Floresta Aleatória)

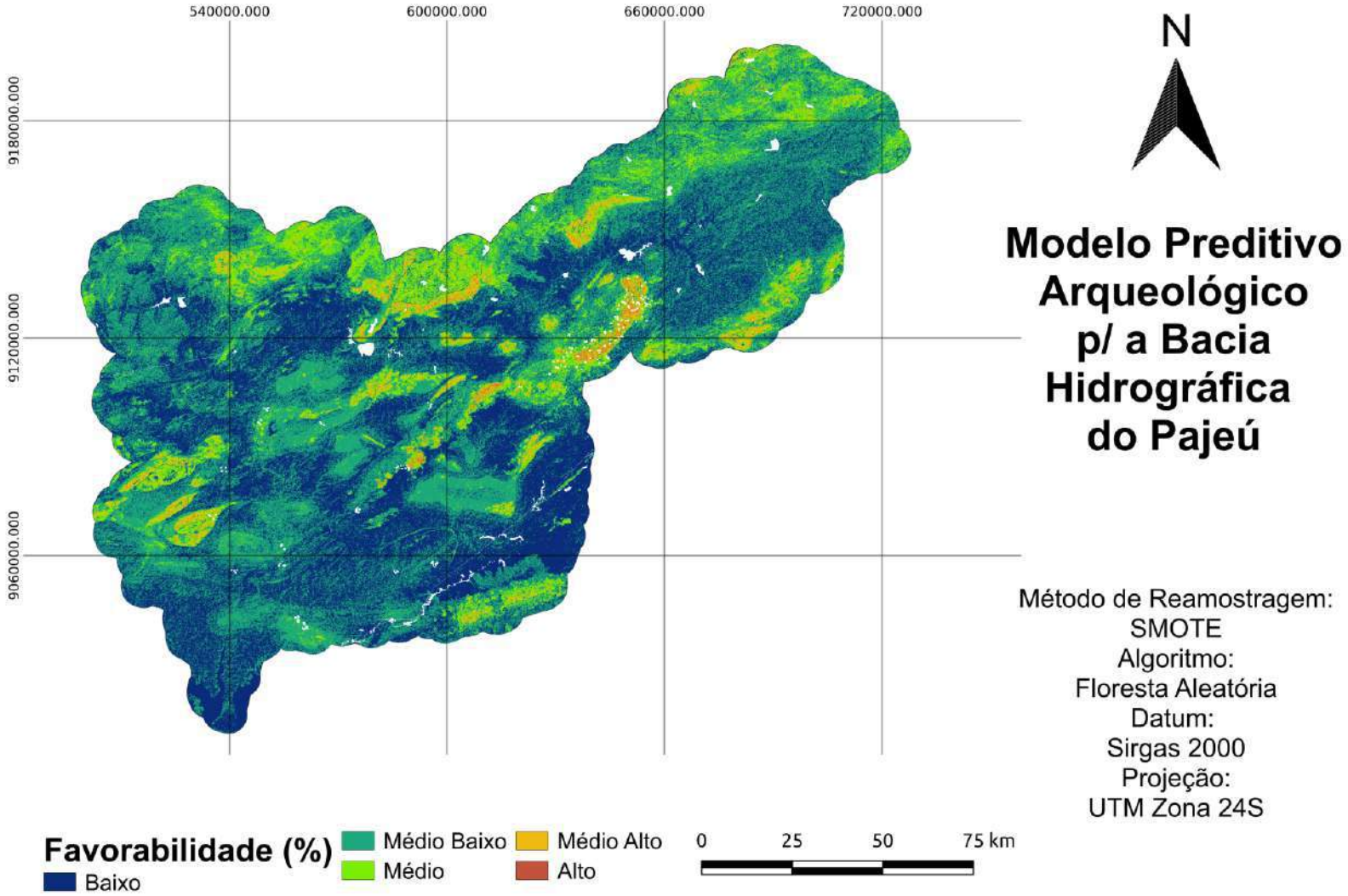


Figura 79: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (Bagging Balanceado)

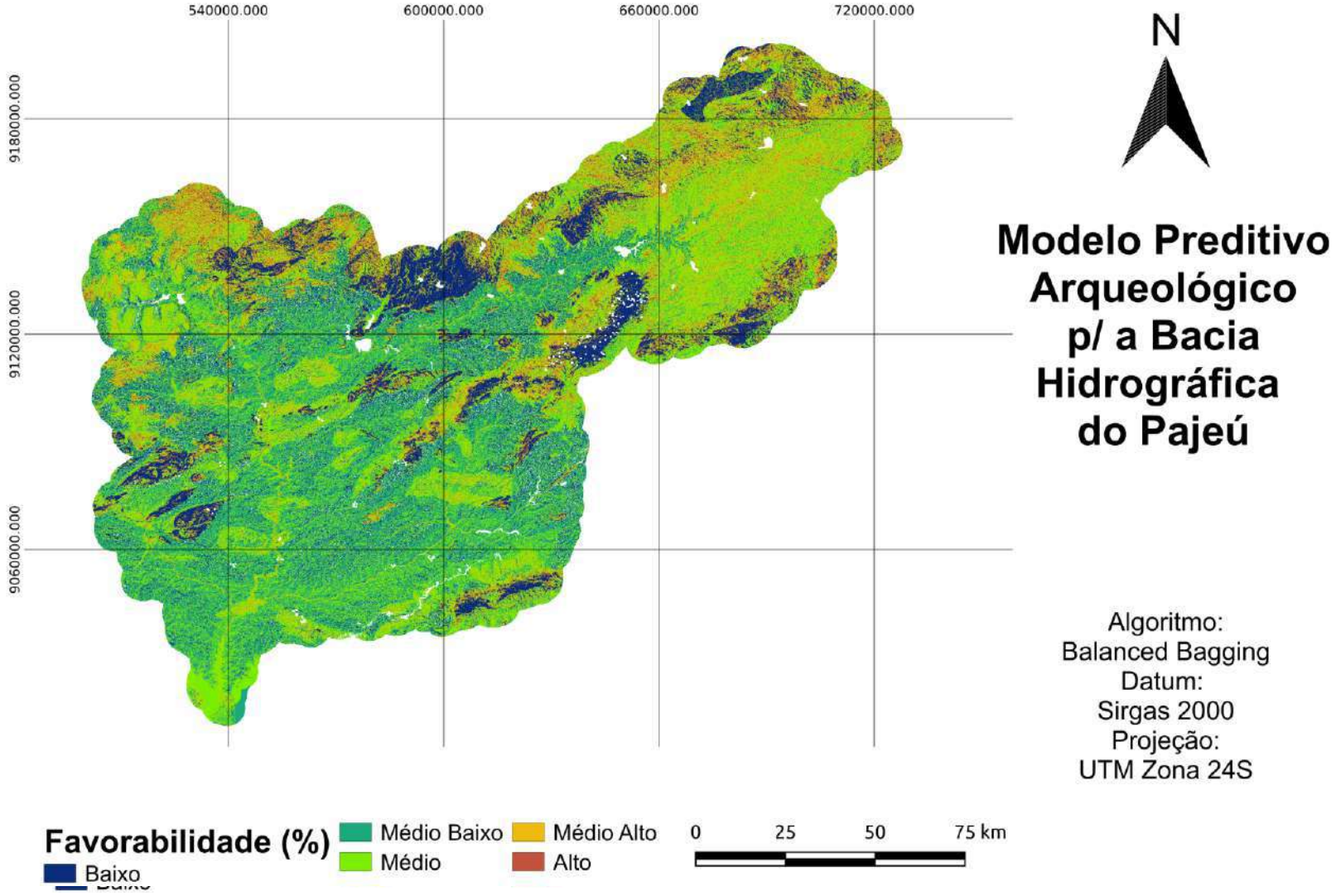
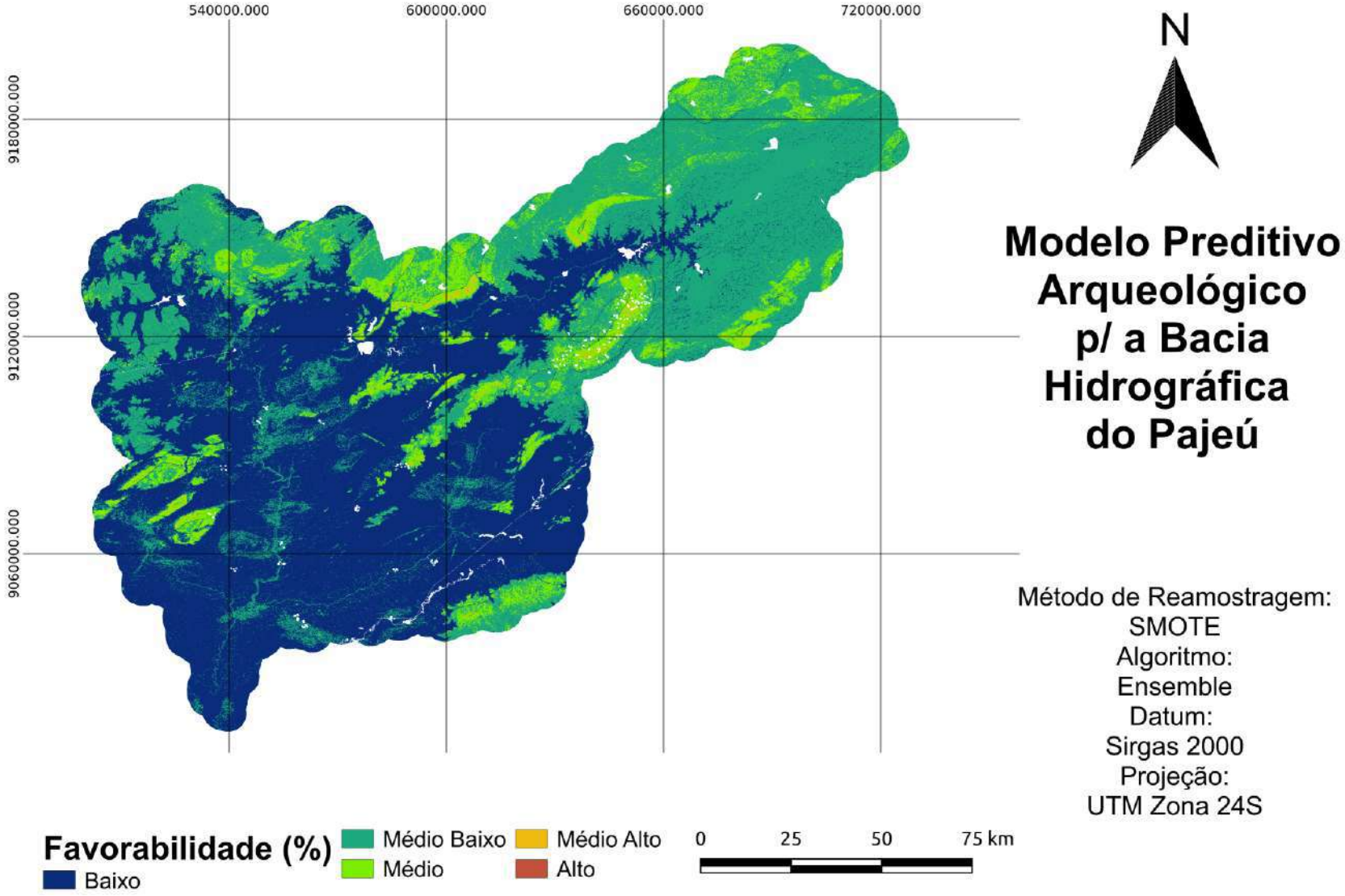


Figura 80: MPA_Pajeu_1.1 com o método de reamostragem SMOTE (Ensemble)



Como é possível observar nos mapas acima (**Figuras 75 a 80**), mesmo com a diminuição significativa do número de variáveis usadas para calcular o modelo, fica evidente que os mesmos locais que já apareciam nas outras versões, continuam a aparecer aqui também. O que chama atenção entre as diferentes versões dos modelos, que neste último há, de modo geral, uma menor quantidade de áreas em amarelo, que representam a favorabilidade média e indicam uma dúvida.

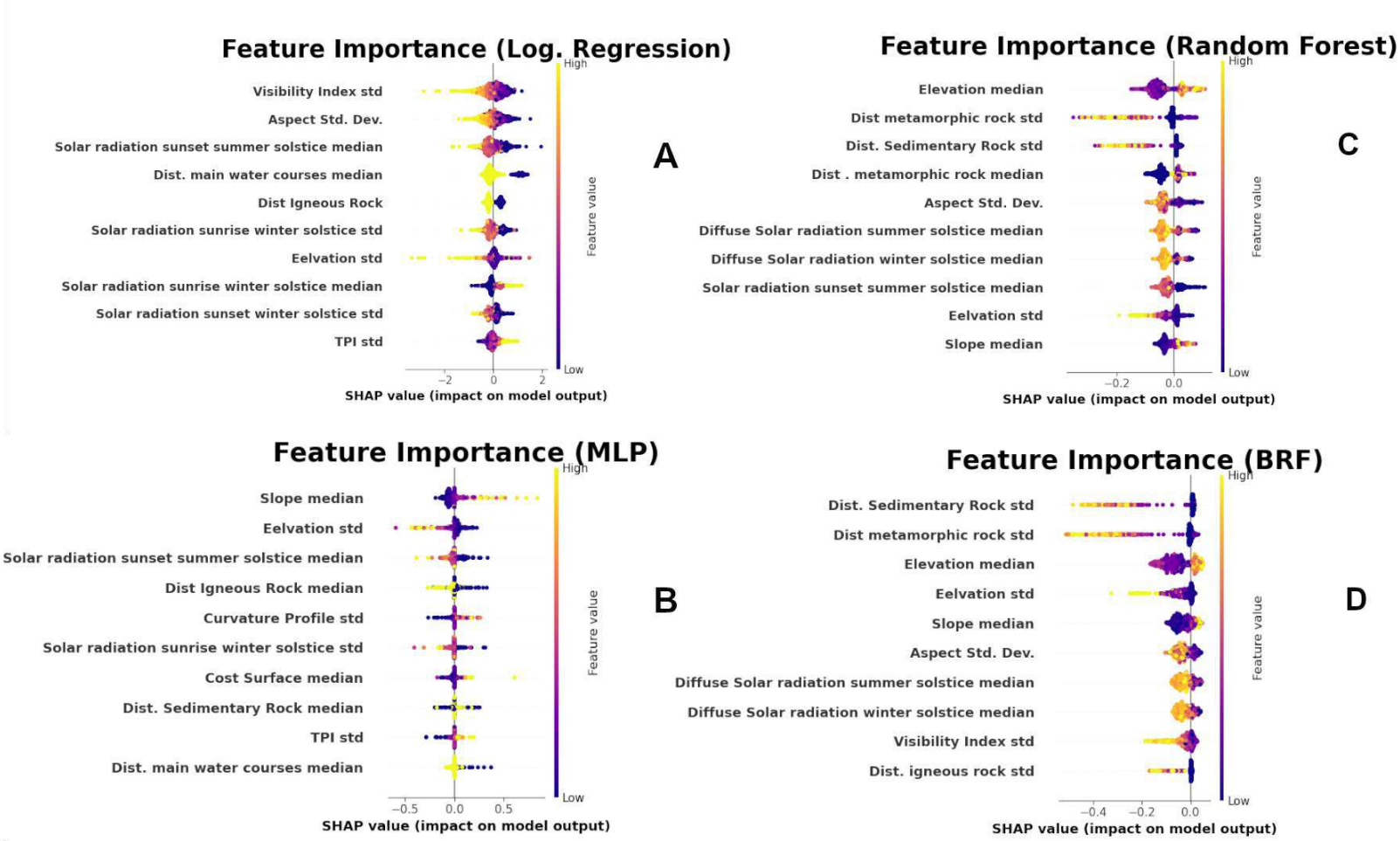
.

..7.6.1 A importância das variáveis

Nesta versão do modelo, com menos variáveis foi possível observar algumas *features* se repetindo em relação aos modelos anteriores, mas também houve uma diferença significativa em alguns pontos. Abaixo é possível observar os gráficas de dispersão *beeswarm*[1] contendo a importância dos modelos que usam os algoritmos de Regressão logística, RNA Perceptron MC, Floresta Aleatória e FAB.

[1] É um tipo de gráfico de dispersão que, neste caso, tem a função não apenas de indicar a média dos valores SHAP de cada uma das variáveis do modelo, mas também de nos informar quanto o valor integral de cada uma dessas características está relacionado com a capacidade de prever localizações como sendo locais de presenças ou ausências. . Em resumo, os valores negativos do SHAP indicam uma maior contribuição para uma ausência, enquanto os positivos indicam uma contribuição para a descoberta de um sítio arqueológico.

Figura 82


Feature Importance (Log. Regression)
Visibility Index std
Aspect Std. Dev.
Solar radiation sunset summer solstice median
Dist. main water courses median
Dist Igneous Rock
Solar radiation sunrise winter solstice std
Eelvation std
Solar radiation sunrise winter solstice median
Solar radiation sunset winter solstice std
TPI std
High
Low
Feature value
−2
0
2
SHAP value (impact on model output)
A
Feature Importance (Random Forest)
Elevation median
Dist metamorphic rock std
Dist. Sedimentary Rock std
Dist . metamorphic rock median
Aspect Std. Dev.
Diffuse Solar radiation summer solstice median
Diffuse Solar radiation winter solstice median
Solar radiation sunset summer solstice median
Eelvation std
Slope median
High
Low
Feature value
−0.2
0.0
SHAP value (impact on model output)
C
Feature Importance (MLP)
Slope median
Eelvation std
Solar radiation sunset summer solstice median
Dist Igneous Rock median
Curvature Profile std
Solar radiation sunrise winter solstice std
Cost Surface median
Dist. Sedimentary Rock median
TPI std
Dist. main water courses median
High
Low
Feature value
−0.5
0.0
0.5
SHAP value (impact on model output)
B
Feature Importance (BRF)
Dist. Sedimentary Rock std
Dist metamorphic rock std
Elevation median
Eelvation std
Slope median
Aspect Std. Dev.
Diffuse Solar radiation summer solstice median
Diffuse Solar radiation winter solstice median
Visibility Index std
Dist. igneous rock std
High
Low
Feature value
−0.4
−0.2
0.0
SHAP value (impact on model output)
D

Pode-se observar na figura acima (**Figura 81)** que as variáveis que se relacionam com a topografia do terreno como aspecto, elevação e declividade são recorrentes na maioria dos diferentes algoritmos calculados. Estes resultados vêm corroborar outros estudos (BANERJEE; SRIVASTAVA, 2014; JOUTEAU et al., 2019; SOUZA, 2016) que mostram que estas variáveis são indicadores muito bons de sítios arqueológicos de arte rupestre.

Olhando mais especificamente, podemos observar que nos gráficos A e D, referentes à regressão logística, e a FAB, o índice de visibilidade aparece como uma característica significativa, o que mostra a importância da inclusão de aspectos socioambientais no MPA. Pode-se perceber que em todos os modelos o baixo valor de radiação solar está diretamente relacionada à presença de sítios. Quando olhamos para as feições geológicas, não é possível observar uma distinção quanto ao tipo de rocha de suporte para a presença de sítios arqueológicos, todos se destacando quase igualmente, com ligeiro destaque para as rochas sedimentares.

..7.6.2 Avaliando estatisticamente o MPA_Pajeu_1.1

Abaixo é possível ver como ficou a avaliação estatística do modelo. Quando olhamos para o resultado geral, podemos observar que o modelo apresenta bom desempenho na previsão correta das classes positivas, como pode ser visto pelos valores de recall. A precisão também apresenta bom desempenho em geral, exceto nos modelos RL e árvore de Decisão. Mesmo sabendo que não tem a mesma relevância que o recall já que o número de falsos positivos não impacta o objetivo final da pesquisa, que é encontrar sítios arqueológicos, podemos observar valores próximos a 50%, dando mais confiança ao modelo .

Os resultados estatísticos parecem promissores, mas vale ressaltar que estamos trabalhando com uma base de dados muito restrita, muito uniforme quando se considera o contexto arqueológico em que cada sítio está inserido, e isto pode contribuir para estes elevados valores preditivos na nossa avaliação estatística.

Tabela 15: Métricas estatísticas de avaliação para os modelos do MPA_Pajeu_1.1.

Métodos de Reamostragem SMOTE - sobreamostragem

ALGORITMOS	Acurácia	Preisão	Recall	F2
FLORESTA ALEATÓRIA	0.95	0.78	0.75	0.76
REGRESSÃO LOGÍSTICA	0.88	0.47	0.75	0.67
RNA MC PERCEPTRON	0.94	0.75	0.75	0.75
ÁRVORE DE DECISÃO	0.94	0.68	0.87	0.83
ENSEMBLE	0.94	0.77	0.71	0.72

Método de reamostragem - Subamostragem

ALGORITMOS	Acurácia	Preisão	Recall	F2
BAGGING BALANCEADO	0.92	0.6	0.87	0.8
FLORERSTA ALEATÓRIA BALANCEADA	0.91	0.57	0.83	0.76

8. CONSIDERAÇÕES FINAIS

Esta pesquisa partiu de um questionamento central que visava responder uma pergunta acerca dos padrões ambientais e sociais que podem estar relacionados aos sítios de grafismos rupestres e para isso foram criados diversos modelos preditivos arqueológicos, que visavam encontrar através de uma base de dados conhecida de sítios arqueológicos presentes na Bacia Hidrográfica do Pajeú, onde poderiam ter áreas propícias para o achado de novos vestígios rupestres e, consequentemente, encontrar elementos que podem ter levado a um indivíduo ou grupo a criar representações gráficas nas rochas.

Diante do que foi posto como problemática desta pesquisa, também formulou-se uma hipótese que dizia que os sítios arqueológicos estavam sim, ligados a aspectos socioambientais e estritamente ambientais e de acordo com todos os resultados apresentados, tal afirmação foi comprovada, de modo que se encontrou elementos socioambientais característicos relacionados aos locais em que se identificam grafismos rupestres.

Do ponto de vista arqueológico foi observado que alguns elementos tais como a orientação das encostas, a declividade, a elevação e o índice de visibilidade, e elementos da busca por recursos naturais tais como a proximidade à rochas sedimentares, podem ser elementos de grande utilidade na busca por padrões arqueológicos.

Em comparação a outras pesquisas do gênero, a adoção de variáveis de cunho socioambiental, que visaram retirar um pouco do determinismo ambiental atrelado aos MPA, mostrou-se um diferencial, já que os CMC traçados durante o trabalho puderam nos mostrar que boa parte dos sítios arqueológicos presentes no Pajeú não se encontram em áreas de grande densidade de caminhos, seja usando a insolação ou a declividade como superfície de custo.

Este aspecto acima mencionado já foi discutido ao longo do texto, mas cabe aqui ressaltar alguns pontos. Primeiramente temos que ter em mente que trabalhamos com uma base de dados que não representa o completo universo de sítios do Pajeú; segundo, os fatores simbólicos devem ter sido levado em consideração pelos grupos, além da questão comunicadora ; terceiro ponto, é preciso levar em consideração a conservação dos grafismos. É sabido que as pinturas passam por intemperismo químico que as apaga com o tempo e os vestígios restantes, que são preservados até os dias atuais, devem representar apenas uma amostra de um universo

maior, que inclusive poderia estar relacionados a áreas mais próximas de CMC com maior densidade, sobretudo se levarmos em conta a função comunicadora destes.

É importante frisar que há outra inúmeras características que poderiam ser aqui inseridas, porém ou se perderam com o tempo, ou estão dentro de uma conotação mais individual relativa a seus autores, Essa variável nunca será possível ser inserido no modelo, pois ser formada por elementos subjetivos, os quais só poderiam ser acessados pelos próprios criadores das artes rupestres do semiárido nordestino. Esta análise retorna ao que Van Havre (2015) já discutia, pois certamente já havia interações entre os grupos, independentes dos registros rupestres.

É claro que esta tese não tem o interesse de responder a todas as perguntas referentes aos estudos dos grafismos rupestres do nordeste do Brasil, mas ela coloca a região do Pajeú no mapa. Esta área que antes apresentava trabalhos pontuais, agora tem um estudo mais sistemático realizado em seu contexto, o que pode levar a um entendimento maior sobre a dispersão dos estilos pictóricos e consequentemente dos grupos humanos que por ali passaram.

Além disso, este trabalho proporcionou o uso de métodos e técnicas advindos de um ramo da Computação, que é a IA e mais precisamente o AM. Estes foram essenciais para a construção do MPA, dando respostas mais graduadas aos modelos e permitindo que as variáveis pudessem ser analisadas de maneira menos enviesada, já que aqui toda a importância das variáveis é calculada a longo do processo de execução do modelo, sem que sejam impostos pesos pelo pesquisador.

Os resultados apresentados podem ser de grande serventia, não só para o uso na bacia hidrográfica estudada, mas ao longo de toda uma região semiárida nordestina, auxiliando pesquisadores em trabalhos acadêmicos e de contrato — um dos diferenciais dos MPA é dar esse suporte científico para trabalhos de Estudos de Impacto Ambiental, de modo a torná-los mais céleres e mesmo assim mantendo a confiança científica.

Fica claro também que os modelos criados ainda precisam passar por um longo processo de aprimoramento que só serão atingidos com a aquisição de mais informações em mais idas a campo. O *MnModel* que foi apresentado no início do texto, é uma prova disso, com uma longa trajetória de condução do projeto e que vem colhendo os frutos, diminuindo os gastos do setor público.

O uso para buscar padrões na diferenciação entre o gravar e o pintar pode também ser explorado com a aquisição desses novos dados, já que neste trabalho a redução da base de dados mostrou-se problemática em termos estatísticos, mas os mapas indicam uma localização um pouco distinta para as diferentes técnicas gráficas. Com a evolução do uso de modelos, a inserção de dados e a própria utilização de outras técnicas, podem levar até a um entendimento mais abrangente sobre esses diferentes modos de registrar nas rochas.

A ampliação do trabalho não passa só pela aquisição de novos dados, mas também pelo uso de novas metodologias. O trabalho deixa a porta aberta para uma ampliação de sua área de abrangência, de modo a compreender os contextos dos locais circunvizinhos e quem sabe melhor compreender às dinâmicas populacionais dos grupos pré-historicos que habitaram os sertões nordestinos.

A ampliação no uso de variáveis sociais tais como a acessibilidade e o índice de visibilidade, podem ser explorados de outras formas, tais como na construção de redes com pontos de origem/destino com distâncias entre eles ainda mais reduzidas, diminuindo um possível erro associado ao se gerar caminhos ótimos.

Além disso, novos algoritmos e métodos são desenvolvidos pela IA frequentemente, portanto estar ambientado às novas inovações tecnológicas pode ser um ponto fundamental para o desenvolvimento de novos produtos, ainda mais qualificados no futuro.

Por fim, é importante reiterar que este trabalho além de não ter a intenção de buscar as respostas para todos os problemas referentes aos estudos espaciais dos grafismos rupestres, também teve a responsabilidade de mostrar que o uso de determinadas tecnologias não tem o interesse de substituir o arqueólogo e muito menos podem ser vistas como “poções mágicas”. A IA é uma técnica que exige um operador que carregará seus vieses, seja na montagem, com a escolha das variáveis e métodos, ou na interpretação de seus produtos.

REFERÊNCIAS

ABU-MOSTAFA, Y. S.; MAGDON-ESMAIL, M.; LIN, H.-T. **Learning from Data: A short course**. 1ª ed. Pasadena: AML Book, 2012.

AGUIAR, A. A Tradição agreste: Estudo sobre a arte rupestre em Pernambuco. **Clio Arqueológica**, v. 1, p. 7–98, 1986.

ALBUQUERQUE, M. Cultivadores pré-históricos no semi-árido:aspectos paleoambientais. **Clio Arqueológica**, v. 4, p. 117–118, 1987.

ANA. **Regiões Hidrográficas**. Disponível em: <https://dadosabertos.ana.gov.br/datasets/b78ea64219b9498c8125cdef390715b7/explore>. Acesso em: 24 ago. 2020.

ANICHINI, F.; WRIGHT, H.; VILA, L. **Talking about the revolution. Innovation in communication within the ARCHAIDE project**. III Congreso Internacional de Buenas Prácticas en Patrimonio Mundial. **Anais**...Mahón: Universidad Complutense de Madrid, 2018

ANNA NILSSON. **Predicting the Archaeological Landscape Archeological Density Estimation around the Ostlänken railroad corridor**. Uppsala: Monografia (Bacharelado em Ciência e Tecnologia) Departamento de Tecnologia da Informação, Universidade de Uppsala, Uppsala, 2016.

ARRUTI, J. M. A. Estudos Históricos. **Morte e vida do nordeste indígena: a emergência étnica como fenômeno regional**, v. 8, n. 15, p. 57–94, 1995.

AZEVEDO, R. L. et al. ESR dating of megafauna enamel teeth from Lagoa Uri de Cima Archaeological Site (Pernambuco, Northeastern Brazil). **Quaternary International**, v. 556, p. 33–48, 2019.

BANERJEE, R.; SRIVASTAVA, K. Remote Sensing Based Identification of Painted Rock Shelter Sites: Appraisal Using Advanced Wide Field Sensor, Neural Network and Field Observations. In: SRIVASTAVA, P. K.; AL., E. (Eds.). . **Remote Sensing Applications in Environmental Research**. [s.l.] SpringerInernational, 2014. p. 195–212.

BANERJEE, R.; SRIVASTAVA, P.; PIKE, A. W. G. Identification of Painted Rock-Shelter

Sites Using GIS Integrated with a Decision Support System and Fuzzy Logic. **ISPRS International Journal of Geo-Information**, v. 7, p. 326, 12 ago. 2018.

BARBOSA, R.; MUTZENBERG, D.; CISNEIROS, D. Perfil Gráfico das Pinturas Rupestres Pré-Históricas do Vale Do Moxotó e Quadrante Nordeste da Bacia Hidrográfica do Pajeú – Pernambuco, Brasil. **Revista Noctua**, v. 1, p. 97–105, 2016.

BARBOSA, R. N. **Perfil gráfico das pinturas rupestres pré-históricas do Vale do Moxotó e Quadrante Nordeste da bacia hidrográfica do Pajeú - Pernambuco, Brasil**. [s.l.] Tese (Doutorado em Arqueologia) Programa de Pós-Graduação em Arqueologia, Universidade Federal de Pernambuco. Recife - PE, 2013.

BARCELO, J. Artificial Intelligence. In: VARELA, S. L. L. (Ed.). . **The Encyclopedia of Archaeological Sciences**. Newark: JohnWiley & Sons, 2018. p. 1–6.

BARCELÓ, J. A. The birth and historical development of computational intelligence applications in Archaeology. **Archeologia e Calcolatori**, v. 20, p. 95–109, 2009.

BARCELÓ, J. A. **Computational Intelligence in Archaeology. State of the Art**. (B. Frischer, J. W. Crawford, D. Koller, Eds.)Making History Interactive Proceedings of the 37th International Conference. **Anais**...Wiliansburg: 2010

BARCELÓ, J. A.; DEL CASTILLO, F. Simulating the Past for Understanding the Present. A Critical Review. In: BARCELÓ, J. A.; DEL CASTILLO, F. (Eds.). . **Simulating Prehistoric and Ancient Worlds**. Cham: Springer International Publishing, 2016. p. 1–140.

BAXTER, M. **Notes on Quantitative Archaeology and R**. 1ª ed. [s.l.] [s.n], 2015.

BEHLING, H. et al. Late Quaternary vegetational and climate dynamics in northeastern Brazil, inferences from marine core GeoB 3104-1. **Quaternary Science Reviews**, v. 19, n. 10, p. 981–994, 2000.

BINFORD, L. R. A Consideration of Archaeological Research Design. **American Antiquity**, v. 29, n. 4, p. 425–441, 1964.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [s.l.] Springer New York, 2011.

BORRA, S.; THANKI, D.; DEY, N. **Satellite Image Analysis: Clustering and**

Classification. Cham: SpringerBriefs in Applied Sciences and Technology, 2019.

BRANDI, P. **Memórias da Eletricidade**. Disponível em: <https://www.memoriadaeletricidade.com.br/artigos/40467/ha-73-anos-a-chesf-comecava-sua-caminhada>. Acesso em: 20 ago. 2021.

BRANDSEN, A. **Digging in documents: using text mining to access the hidden knowledge in Dutch archaeological excavation reports**. [s.l.] Tese(Doutorado em Arqueologia) Departamento de Arqueologia - Universidade de Leiden. Leiden, Países Baixos, 21 fev. 2022.

BRASIL. **Mapa de Geodiversidade do Brasil ao Milionésimo**. Disponível em: <https://geosgb.cprm.gov.br/>. Acesso em: 2 jan. 2020.

BROWNLEE, J. **Imbalanced Classification with Python: Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning**. 1.2 ed. [s.l.] Jason Brownlee, 2020.

BURKOV, A. **The Hundred-Page Machine Learning Book**. [s.l.] Andriy Burkov, 2019.

CAMARGO, C. S. DE. **A água e seus fluxos: ação coletiva, conflitos territoriais e povos indígenas na transposição do Rio São Francisco**. [s.l.] Tese (Doutorado em Ciências Sociais) Universidade Estadual de Campinas, Campinas, 2017.

CARLSON, D. L. **Quantitative Methods in Archaeology Using R**. Cambridge: Cambridge University Press, 2017.

CBHSF. **Angiquinho, a primeira usina hidrelétrica do nordeste**. Disponível em: <https://cbhsaofrancisco.org.br/noticias/cultura_blog/angiquinho-a-primeira-usina-hidreletrica-do-nordeste/>. Acesso em: 20 ago. 2021.

CHAGAS JUNIOR, J. N. **ARQUEOLOGIA ESPACIAL NO SERIDÓ POTIGUAR: análise e interpretação arqueológica do território na bacia hidrográfica do rio Carnaúba**. [s.l.] Dissertação (Mestrado em Arqueologia) Programa de Pós-Graduação em Arqueologia, Universidade Federal de Pernambuco, Recife-PE. 140p., 2017.

CHAWLA, N. et al. SMOTE: Synthetic Minority Over-sampling Technique. **J. Artif. Intell. Res. (JAIR)**, v. 16, p. 321–357, 1 jun. 2002.

CHEN, C.; LIAW, A.; BREIMAN, L. **Using Random Forest to Learn Imbalanced Data**.

Berkeley: ’Berkeley University of California, 2004.

CHOHFI, R. E. et al. **Sensoriamento remoto em levantamento arqueológico: teste de detecção no Vale do Paraíba do Sul. Brasil**. IV Latin American Remote Sensing Symposium. **Anais**...Bariloche: 1989

CLARK, G. **Prehistoric Europe: the economic basis**. Londres: Methuen, 1952.

CLARKE, D. L. **Analytical Archaeology**. Londres: Methuen, 1968.

CLARKE, D. L. **Spatial archaeology**. Cambridge: Academic Press, 1977.

CODEVASF. **Sistema Itaparica**. Disponível em: <https://www.codevasf.gov.br/linhas-de-negocio/irrigacao/sistema-itaparica>. Acesso em: 20 ago. 2021.

CONOLLY, J.; LAKE, M. **Geographical Information Systems in Archaeology**. Madri: Bellatera, 2009.

CORMEN, T. H. et al. **Introduction to algorithms**. Cambridge: The MIT Press, 2001.

CORRÊA, A. et al. The Semi-arid Domain of the Northeast of Brazil. In: SALGADO, A. A. R.; SANTOS, L. J. C.; PAISANI, J. . (Eds.). . Nova Iorque: Springer Inernational, 2019. p. 119–150.

COSTA, F. A. P. DA. **Anais Pernambucanos**. 10v. ed. Recife: Arquivo Público Estadual, 1952.

CUNHA, E. **Os sertões**. 1ª ed. Rio de Janeiro: Fundação Darcy Ribeiro, 2013.

COWLEY, D. et al. Expedite and upscale large area archaeological survey in the age of big data and machine learning. v. 112, p. 15–17, 1 jan. 2021.

CRABTREE, S. A. et al. Landscape rules predict optimal superhighways for the first peopling of Sahul. **Nature Human Behaviour**, v. 5, n. 10, p. 1303–1313, 2021.

DANTAS, Z.; GONZAGA, L. **Riacho do Navio**Rio de JaneiroUniversal Music, , 1955.

DAYET, L.; WURZ, S.; DANIEL, F. Ochre Resources, Behavioural Complexity and Regional Patterns in the Howiesons Poort: New Insights From Klasies River Main Site, South Africa. **Journal of African Archaeology**, v. 15, p. 20–41, 7 dez. 2017.

DEFAZIO, A.; BACH, F.; LACOSTE-JULIEN, S. **SAGA: A fast incremental gradient**

method with support for non-strongly convex composite objectives. Advances in neural information processing systems. **Anais**...2014

DIJKSTRA, E. W. A note on two problems in connexion with graphs. **Numerische Mathematik**, v. 1, n. 1, p. 269–271, 1959.

DINCAUZE, D. F. **Environmental Archaeology: Principles and Practice**. [s.l.] Cambridge University Press, 2000.

DIWAN, G. A. GIS-Based Comparative Archaeological Predictive Models: A First Application to Iron Age Sites in the Bekaa (Lebanon). **Mediterranean Archaeology and Archaeometry**, v. 20, n. 2, p. 143–158, 2020.

DORAN, J. Knowledge Representation for Archaeological Inference. In: MICHIE, E. W. E. AND D. (Ed.). . **Machine Intelligence 8**. [s.l.] Ellis Horwood Ltd, 1977. p. 433–454.

DOUGLASS, M.; LIN, S.; CHODORONEK, M. The Application of 3D Photogrammetry for In-Field Documentation of Archaeological Features. **Advances in Archaeological Practice**, v. 3, p. 136–152, 1 maio 2015.

DRIES, M. H. VAN DEN. **Archaeology and the application of artificial intelligence : case-studies on use-wear analysis of prehistoric flint tools**. [s.l.] University of Leiden, 1998.

DUMPALA, S. H.; CHAKRABORTY, R.; KOPPARAPU, S. K. **A Novel Data Representation for Effective Learning in Class Imbalanced Scenarios**. [s.l: s.n.].

ETCHEVARNE, C. Ambiente e ocupação humana em uma região do Sub-médio São Francisco, Bahia. **Clio Arqueológica**, v. 15, p. 61–88, 2002.

FERNÁNDEZ, A. et al. **Learning from Imbalanced Data Sets**. Cham: Springer Inernational, 2018.

FIORE, D. Las dimensiones espaciales del arte parietal. Estado actual de las investigaciones desde el Cono Sur y propuesta de síntesis metodológica. In: **Imagenes Rupestres: Lugares y Regiones.** Buenos Aires: Universidad Nacional de Rosario. Facultad de Humanidades y Arte, 2018. p. 51-62.

FITZ, P. R. **Geoprocessamento sem complicação**. 1. ed. São Paulo: Oficina de Textos, 2008.

FLORES, F. C. et al. Computer Algorithm for Archaeological Projectile Points Automatic

Classification. **J. Comput. Cult. Herit.**, v. 12, n. 3, 2019.

FONSECA, J. A. DA. Modelagem espacial de sítios arqueológicos nas Serras de Carajás: as inferências dos modelos preditivos. In: MAGALHÃES, M. P. (Ed.). . **Humanidade e a Amazônia: 11 mil anos de evolução histórica em Carajás**. Belém: Museu Paraense Emílio Goeldi, 2018a. p. 211–231.

FONSECA, J. A. DA. Padrões de distribuição espacial e modelos preditivos: os sítios arqueológicos no baixo curso dos rios Nhamundá-Trombetas. **Boletim do Museu Paraense Emílio Goeldi Ciências Humanas**, v. 13, n. 2, p. 353–376, 2018b.

FORESTGIS. **Download GIS (Banco de Dados de Links)**. Disponível em: <https://forest-gis.com/download-gis-base-de-dados/>. Acesso em: 2 jan. 2020.

FORTE, M.; CAMPANA, S. **Digital Methods and Remote Sensing in Archaeology**. Cham: Springer Inernational, 2016.

FRANÇA, L. M. D. A. **Impacto das ações antrópicas e do clima no uso e ocupação do solo na Bacia Hidrográfica do Rio Pajeú**. [s.l.] Dissertação (Mestrado em Meio Ambiente) Programa de Pós-Graduação em Desenvolvimento e Meio Ambiente - Universidade Federal de Pernambuco, Recife, 2017.

GALLOTTI, R. et al. GIS and Intra-Site Spatial Analyses: An Integrated Approach for Recording and Analyzing the Fossil Deposits at Casablanca Prehistoric Sites (Morocco). **J. Geographic Information System**, v. 3, p. 373–381, 1 jan. 2011.

GALVÃO, D. **Evolução paleoambiental da lagoa do Puiú, Municipio de Ibimirim-PE**. [s.l.] Dissertação (Mestrado em Geografia) Programa de Pós-Graduação em Geografia - Univiersidade Federal de Pernambuco, Recife, 2012.

GÉRON, A. **Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems**. 2nd. ed. Sebastopol: O'Reilly Media, Inc., 2019.

GILLINGS, M. Mapping invisibility: GIS approaches to the analysis of hiding and seclusion. **Journal of Archaeological Science**, v. 62, p. 1–14, 2015.

GILLINGS, M.; HACIGUZELLER, P.; LOCK, G. Archaeology and spatial analysis. In:

GILLINGS, M.; HACIGÜZELLER, P.; LOCK, G. (Eds.). . **Archaeological Spatial Analysis: A Methodological Guide**. Londres: Routledge, 2019.

GLOCK, A. Cultural Bias in the Archaeology of Palestine. **Journal of Palestine Studies**, v. 24, n. 2, p. 48–59, 1995.

GOVERNO DE PERNAMBUCO. **PE3D**.

GREGORY, K. J.; LEWIN, J. A hierarchical framework for concepts in physical geography. **Progress in Physical Geography: Earth and Environment**, v. 42, n. 6, p. 721–738, 3 set. 2018.

GUIDON, N. A arte pré-histórica da área arqueológica de São Raimundo Nonato: síntese de dez anos de pesquisas. **Clio Arqueológica**, v. 2, p. 3–80, 1985.

GUPTA, V. **Kaggle**. Disponível em: <https://www.kaggle.com/code/jinxzed/interpreting-logistic-regression-using-shap/notebook>.

HAYASHI, C. What is Data Science? Fundamental Concepts and a Heuristic Example. In: HAYASHI, C. et al. (Eds.). . **Data Science, Classification, and Related Methods**. Tóquio: [s.n.]. p. 40–51.

HE, H. et al. **ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning**. IEEE International Joint Conference on Neural Networks. **Anais**...Pequim: 8 jul. 2008

HERZOG, I. Theory and practice of cost functions. **Fusion of Cultures. Proceedings of the 38th Annual Conference on Computer Applications and Quantitative (Methods in Archaeology**, p. 375–382, 1 jan. 2013.

HERZOG, I. Least-cost Paths – Some Methodological Issues. **Internet Archaeology**, 1 jun. 2014.

HOBBS, E. **MnModel Phase 4**. Saint Paul: Minessota Department of Transportation, 2019.

HOFIERKA, J.; RI, M. The solar radiation model for Open source GIS: Implementation and applications. **Proceedings of the Open Source GIS-GRASS Users Conference**, 24 out. 2002.

HUFF, D. **How to lie with statistics**. 1ª ed. New York: W. W. Norton & Company, 1956.

HUGGETT, J. Challenging Digital Archaeology. **Open Archaeology**, v. 1, 3 jan. 2015.

HUGGETT, J. The Apparatus of Digital Archaeology. **Internet Archaeology**, v. 44, 1 jan. 2017.

HUGGETT, J. Is Big Digital Data Different? Towards a New Archaeological Paradigm. **Journal of Field Archaeology**, v. 45, n. sup1, p. S8–S17, 20 fev. 2020.

HUGGETT, J.; REILLY, P.; LOCK, G. Whither Digital Archaeological Knowledge? The Challenge of Unstable Futures. **Journal of Computer Applications in Archaeology**, v. 1, p. 42–54, 23 maio 2018.

IBGE. **Semiárido Brasileiro**. Disponível em: <https://www.ibge.gov.br/geociencias/cartas-e-mapas/mapas-regionais/15974-semiarido-brasileiro.html?=&t=o-que-e>. Acesso em: 22 ago. 2021.

IPHAN. **sitios_arqueologicos_cnsa**. Disponível em: <http://portal.iphan.gov.br/uploads/ckfinder/arquivos/Legendas_arquivos_shapefile_sitios_arqueologicos_cnsa_shp.pdf>.

ISNARDIS, A.; LINKE, V. Pedras pintadas, paisagens construídas: a integração de elementos culturalmente arquitetados na transformação e manutenção da paisagem. **Revista de Arqueologia**, v. 23, n. 1, p. 42–59, 2010.

JAMES, G. et al. **An Introduction to Statistical Learning: with Applications in R**. [s.l.] Springer New York, 2014.

JOUTEAU, A. et al. Choosing rock art locations: Geological parameters and social behaviours. The example of Cussac Cave (Dordogne, France). **Journal of Archaeological Science**, v. 105, p. 81–96, 2019.

KELLEHER, J. D.; MACNAMEE, B.; D'ARCY, A. **Fundamentals of Machine Learning for Predictive Data Analytics**. Cambridge: MIT Press, 2015.

KEMPF, M. The application of GIS and satellite imagery in archaeological land-use reconstruction: A predictive model? **Journal of Archaeological Science: Reports**, v. 25, p. 116–128, 2019.

KERR, S. The future of archaeology, interdisciplinarity and global challenges. **Antiquity**, v.

94, p. 1–12, 1 set. 2020.

KIPNIS, R. **O uso de modelos preditivos para diagnosticar recursos arqueológicos em áreas a serem afetadas por empreendimentos de impacto ambiental**. (S. B. CALDARELLI, Ed.)Atas do Simpósio sobre Política Nacional do Meio Ambiente e Patrimônio Cultural: Repercussões dos Dez Anos da Resolução CONAMA nº 001/86 sobre a Pesquisa e a Gestão dos Recursos Culturais no Brasil. **Anais**...Goiânia: 1997

KLEHM, C. et al. Toward archaeological predictive modeling in the Bosutswe region of Botswana: Utilizing multispectral satellite imagery to conceptualize ancient landscapes. **Journal of Anthropological Archaeology**, v. 54, p. 68–83, 2019.

KOOPS, B.-J.; GALIČ, M. Conceptualising Space and Place: Lessons from Geography for the Debate on Privacy in Public. In: TIMAN, T.; NEWELL, B.; KOOPS, B.-J. (Eds.). . **Privacy in public space**. Cheltenham: Edward Elgar Publishing, 2017. p. 19–46.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. New York, NY, USA: Springer New York, 2013.

LANGMUIR, E. **Mountaincraft and Leadership**. 4ª ed. Leicester: MLTE, 2013.

LEMAIÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. **Journal of Machine Learning Research**, v. 18, n. 17, p. 1–5, 2017.

LESMEISTER, C. **Mastering Machine Learning with R**. Birminghan: Packt Publishing, 2015.

LIMA, J. P. DE; MEDEIROS, R. P.; CASTRO, V. M. C. DE. Grupos Ceramistas da Ilha de Sorobabel, Itacuruba-PE: Continuidade e Mudança Tecnológica. **Revista Noctua**, v. 1, p. 43–65, 2016.

LOCK, G.; POUNCETT, J. Spatial thinking in archaeology: Is GIS the answer? **Journal of Archaeological Science**, v. 84, 1 jun. 2017.

LOURDEAU, A.; PAGLI, M. Caracterização tecnológica da indústria lítica de Lagoa Uri de Cima. **Fumdhamentos**, v. 10, p. 103–131, 1 jan. 2013.

LUNDBERG, S.; LEE, S.-I. **A Unified Approach to Interpreting Model Predictions**. [s.l:

s.n.].

MADEIRA, L. **Padrões de abrigos com registros rupestres no Nordeste do Brasil**. [s.l.] Monografia (Bacharelado em Arqueologia) Departamento de Arqueologia Universidade Federal de Pernambuco, Recife, 2016.

MALAPERDAS, G.; ZACHARIAS, N. The habitation Model Trend Calculation (MTC): A new effective tool for predictive modeling in archaeology. **Geo-spatial Information Science**, v. 22, n. 4, p. 314–331, 2 out. 2019.

MARCHETTI, N. et al. NEARCHOS. Networked Archaeological Open Science: Advances in Archaeology Through Field Analytics and Scientific Community Sharing. **Journal of Archaeological Research**, v. 26, n. 4, p. 447–469, 2018.

MARTIN, G. O Povoamente Pré-histórico do Vale do São Francisco. **Clio Arqueológica**, v. 13, p. 9–41, 1998.

MARTIN, G. **Pré-História do Nordeste do Brasil**. Recife: Editora Universitária da UFPE, 2013.

MARTIN, G.; GUIDON, N. A onça e as orantes: uma revisão das classificações tradicionais dos registros rupestres do NE do Brasil. **Clio Arqueológica**, v. 25, n. 1, p. 11–30, 2010.

MARTIN, G.; VIDAL, I. A. Dispersão e difusão das tradições rupestres no Nordeste do Brasil. Vias de ida e volta? **Clio Arqueológica**, v. 29, n. 2, p. 17–30, 2014.

MARWICK, B. et al. OPEN SCIENCE IN ARCHAEOLOGY. **SAA Archaeological Record**, v. 17, p. 8–14, 1 set. 2017.

MATOS, F. DE. **ENTRE SEMELHANÇAS GRÁFICAS E AMBIENTAIS: as recorrências das representações antropomórficas pintadas pré-históricas entre as regiões do Cariri Ocidental-PB, Parque Nacional do Catimbau-PE e Seridó Oriental-RN**. [s.l.] Tese (Doutorado em Arqueologia) Programa de Pós-Graduação em Arqueologia, Universidade Federal de Pernambuco. Recife - PE, 2019.

MAYO, H. et al. **History of Machine Learning**. Disponível em: <https://www.doc.ic.ac.uk/~jce317/history-machine-learning.html>. Acesso em: 10 maio. 2020.

MEDEIROS, R. P. DE; MUTZENBERG, D. Cartografia histórica das relocações indígenas nas ilhas do Submédio São Francisco no período pombalino (1759-1761). **Revista Ultramares**, v. 1, n. 5, p. 1–19, 2014.

MEDEIROS, R. P. DE; MÜTZENBERG, D. S. Cartografia histórica dos povos indígenas em Pernambuco no século XVIII. **Clio Arqueológica**, v. 28, n. 2, p. 1–19, 2013.

MILLMAN, K. J.; AIVAZIS, M. Python for Scientists and Engineers. **Computing in Science & Engineering**, v. 13, n. 2, p. 9–12, 2011.

MISSURA, R. **Bacia do Riacho Pioré-PE, Análise Morfotectônica e Morfoestratigráfica**. [s.l.] Tese (Doutorado em Geografia)Programa de Pós-Graduação em Geografia - Universidade Federal de Pernambuco. Recife, 2013.

MITCHELL, T. M. **Machine Learning**. [s.l.] McGraw-Hill, 1997.

MOROZOV, E. **BIG TECH: A ascenção dos dados e a morte da política**. São Paulo: Ubu Editora, 2018.

MURRIETA-FLORES, P. Understanding human movement through spatial technologies. The role of natural areas of transit in the Late Prehistory of South-western Iberia. **Trabajos de Prehistoria**, v. 69, p. 103–122, 30 jun. 2012.

MÜTZENBERG, D. et al. Sítio Arqueológico Lagoa Uri de Cima: cronoestratigrafia de eventos paleoambientais no semiárido nordestino. **Fumdhamentos**, v. 10, p. 51–67, 2013.

NASA. **Earthdata**. Disponível em: <https://search.earthdata.nasa.gov/search>. Acesso em: 25 ago. 2020.

NASH, G.; CHIPPINDALE, C. **The Figured Landscapes of Rock-Art: Looking at Pictures in Place**. [s.l.] Cambridge University Press, 2004.

NATURE. **Time to recognize authorship of open data**. Disponível em: <https://www.nature.com/articles/d41586-022-00921-x>.

NILSSON STUTZ, L. A Future for Archaeology: In Defense of an Intellectually Engaged, Collaborative and Confident Archaeology. **Norwegian Archaeological Review**, v. 51, n. 1–2, p. 48–56, 3 jul. 2018.

NOVIELLO, M. et al. Investigating the distribution of archaeological sites: Multiparametric

vs probability models and potentials for remote sensing data. **Applied Geography**, v. 95, p. 34–44, 2018.

NUMPY-TEAM. **About us**. Disponível em: <https://numpy.org/about/>. Acesso em: 27 ago. 2021.

O. H. MYERS. Some Applications of Statistics to Archaeology. **Proceedings of the Prehistoric Society**, v. 16, p. 200–201, 1950.

O'NEIL, C. **Algoritmos de Destruição em Massa**. [s.l.] Editora Rua de Sabão, 2021.

OLIVEIRA, A. L. O Sítio Arqueológico Alcobaça: o sítio refrência no Vale do Catimbau - Buíque - PE. **Clio Arqueológica**, v. 21, n. 2, p. 05–39, 2006.

OLIVEIRA, E. G. DA S.; SILVA, E. H. Os Usos dos Recursos Naturais pelos Indígenas Pankará na Serra do Arapuá (Carnaubeira da Penha/PE). **Fronteiras: Journal of Social, Technological and Environmental Science**, v. 4, n. 3, p. 155–174, 2015.

OLIVITO, R.; TACCOLA, E.; ALBERTINI, N. Cultural Heritage and Digital Technologies. Theory, methods and tools for the study and dissemination of knowledge in the archaeological practice. In: [s.l: s.n.]. p. 475–494.

ONU, F.; UCHE-NWACHI, E.; CHIGBUNDU, E. The Role of Object-Oriented Programming (OOP) in Modeling of Geographic Information Systems (GIS). **IOSR Journal of Mobile Computing & Application**, v. 03, p. 40–46, 1 abr. 2016.

PACHECO DE OLIVEIRA, J.; FREIRE, C. A. DA R. **A Presença Indígena na Formação do Brasil**. Brasília: MEC/Unesco, 2006.

PEREIRA NETO, M. C.; SILVA, N. M. DA. Relevos residuais (maciços, inselbergues e cristas) como refúgios da biodiversidade no Seridó Potiguar. **REVISTA GEONORTE**, v. 3, n. 4, p. 262–273, 2012.

PEREZ, G. et al. O uso do programa R na Arqueologia Paulista: um estudo de caso dos grupos ceramistas Tupiguarani e Itararé-Taquara. **Revista de Arqueologia**, v. 32, p. 84, 28 jun. 2019.

PESSIS, A.-M. Apresentação gráfica e apresentação social na tradição nordeste de pintura rupestre do Brasil. **Clio Arqueológica**, v. 5, p. 11–38, 1989.

PESSIS, A.-M. Identidade e classificação dos registros gráficos pré-históricos do nordeste do Brasil. **Clio Arqueológica**, v. 8, p. 35–66, 1992.

PESSIS, A.-M.; CISNEIROS, D.; MUTZENBERG, D. Identidades Gráficas nos Registros Rupestres do Parque Nacional Serra da Capivara, Piauí , Brasil. **Fumdhamentos**, v. 15, n. 2, 2018.

PESSIS, A. et al. CARACTERIZAÇÃO DOS SÍTIOS PRÉ-HISTÓRICOS COM GRAFISMOS RUPESTRES NO ESTADO DE PERNAMBUCO, BRASIL. **Revista Clio**, v. 32, p. 165–188, 28 ago. 2018.

PETERS, E.; OLIVEIRA, E. V. Marcas em fósseis de megafauna em Lagoa da Pedra, Salgueiro-PE. **Clio Arqueológica**, v. 34, n. 3, p. 45–60, 2019.

PIETRONI, E. From Remote to Embodied Sensing: New Perspectives for Virtual Museums and Archaeological Landscape Communication. In: FORTE, M.; CAMPANA, S. (Eds.). . **Digital Methods and Remote Sensing in Archaeology**. Cham: Springer Inernational, 2016. p. 437–474.

PILLAY, P.; O'REGAN, G.; EMMITT, J. A Locational Analysis of Rock Art in the North Island, Aotearoa New Zealand. 8 maio 2020.

PLOG, S.; PLOG, F.; WAIT, W. Decision Making in Modern Surveys. **Advances in Archaeological Method and Theory**, v. 1, p. 383–421, 7 nov. 1978.

POSTGRESQL. **About**. Disponível em: <https://www.postgresql.org/about/>. Acesso em: 27 set. 2021.

PROUS, A.; JUNQUEIRA, P. A.; MALTA, I. M. Arquelogia do Alto Médio São Francisco Região de Januária e Montalvânia. **Revista de Arqueologia**, v. 2, n. 1 SE-Artigo, p. 59–72, 30 jun. 1984.

PUGIN, J. **Locating the rock art of the Maloti- Drakensberg:** [s.l.] University of the Witwatersrand, 2016.

PYPL. **PYPL PopularitY of Programming Language**. Disponível em: <https://pypl.github.io/PYPL.html>. Acesso em: 7 jan. 2021.

PYTHON.ORG. **BeginnerGuide/Overview**. Disponível em:

<https://wiki.python.org/moin/BeginnersGuide/Overview>. Acesso em: 27 ago. 2021.

PYTHON.ORG. **Módulos**. Disponível em: <https://docs.python.org/pt-br/3/tutorial/modules.html>. Acesso em: 27 ago. 2021.

QGIS. **QGIS - A liderança do SIG de código aberto**. Disponível em: <https://qgis.org/pt_BR/site/about/index.html>. Acesso em: 23 ago. 2021.

RAHMATI, O. et al. Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. **Geomorphology**, v. 298, p. 118–137, 2017.

RENFREW, S. F. C.; ZUBROW, E. B. W. **The Ancient Mind: Elements of Cognitive Archaeology**. [s.l.] Cambridge University Press, 1994.

RIBEIRO, A. S. **Estado de conservação das nascentes do alto trecho do Rio Pajeú, Pernambuco, Brasil**. [s.l.] Tese (Doutorado em Geografia) Programa de Pós-Graduação em Geografia - Universidade Federal de Pernambuco, Recife, 2014.

RIBEIRO, E. P. et al. Estimativa dos índices de vegetação na detecção de mudanças ambientais na bacia hidrográfica do rio Pajeú. **Geosul**, v. 31, n. 62, p. 59–62, 2016.

RIBEIRO, L. Repensando a tradição: a variabilidade estilística na arte rupestre do período intermediário de representações no alto-médio rio São Francisco. **Revista do Museu de Arqueologia e Etnologia**, v. 17, p. 127–147, 2007.

ROBRAHN-GONZÁLEZ, E. Arqueologia em perspectivas: 150 anos de prática e reflexão no estudo de nosso passado. **Revista USP**, v. 0, p. 10, 28 fev. 2000.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3ª ed. Harlow: [s.n.]. v. 1132

SANCHIDRIÁN, J. L. **Manual de Arte Prehistórico**. Madri: Ariel Prehistoria, 2001.

SANJUÁN, L. G. **Introducción al Reconocimiento y Análisis Arqueológico del Territorio**. 1. ed. Barcelona: Ariel Prehistoria, 2005.

SANTOS, J. O. DOS et al. Sítios arqueológicos do município de Flores, Estado de Pernambuco. **INTESA**, v. 6, n. 1, p. 67–129, 2012.

SANTOS JUNIOR, V. DOS. **Registros rupestres da área arqueológica de Santana (RN)**.

[s.l.] Dissertação (Mestrado em Arqueologia) Programa de Pós-Graduação em Arqueologia, Universidade Federal de Pernambuco, Recife - PE. 211 p., 2005.

SANTOS, M. **Técnica, Espaço, Tempo: Globalização e meio técnico-científico_ informacional**. São Paulo: Editora da Universidade de São Paulo, 2008.

SARMENTO, G. T. A. **Virgulino cartografado: relações de poder e territorializações do cangaceiro Lampião (1920-1928)**. [s.l.] Dissertação (Mestrado em História) Programa de pó-Graduação em História. Natal, RN, 2019.

SCHAEFER, J. L. **Decisions Set in Stone: Spatial Analyses of Ozark Rock Art Sites, Elements, and Motifs with GIS**. [s.l.] University of Arkansas, 2016.

SCHIFFER, M. Behavioral Archaeology. In: **Encyclopedia of Archaeology**. [s.l: s.n.]. p. 909–919.

SCHIFFER, M. B. **Formation Processes of the Archaeological Record**. Salt Lake City: University of Utah Press, 1996.

SCHMIDT, S.; MARWICK, B. Tool-Driven Revolutions in Archaeological Science. **Journal of Computer Applications in Archaeology**, v. 3, p. 18–32, 28 jan. 2020.

SCIKIT-LEARN.ORG. **Decision Trees**. Disponível em: <https://scikit-learn.org/stable/modules/tree.html>. Acesso em: 30 ago. 2021.

SCIKIT-LEARN.ORG. **ConfusionMatrixDisplay**. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ConfusionMatrixDisplay.html>. Acesso em: 30 maio. 2022.

SILVA-JÚNIOR, L. S. DA. Brejos de altitude, refúgios para os grupos humanos do sertão: o caso da Serra do Arapuá, Floresta - PE. **Clio Arqueológica**, v. 14, p. 237–244, 2000.

SILVA, A. L. C. **Estudo do balanço hídrico na bacia do Rio Pajeú utilizando o modelo SWIM**. [s.l.] Tese (Doutorado em Engenharia Civil) Programa de Pós-Graduação em Engenharia Civil – Universidade Federal de Pernambuco, Recife, 2019.

SILVA, D. G. D. **Reconstrução da dinâmica geomorfológica do semiárido brasileiro no quaternário superior a partir de uma abordagem multiproxy**. [s.l.] Dissertação (Mestrado em Geografia) Programa de Pós-Graduação em Geografia - Universidade Federal de

Pernambuco, Recife, 2013.

SILVA, E. Índios no semiárido nordestino: (re)conhecendo sociodiversidades. **Clio**, v. 35, n. 1, p. 254–272, 2017a.

SILVA, E. Índios no Nordeste: por uma história socioambiental regional. **Cadernos do Ceas**, v. 240, p. 117–136, 2017b.

SIRTOLI, A. et al. ATRIBUTOS DO RELEVO DERIVADOS DE MODELO DIGITAL DE ELEVAÇÃO E SUAS RELAÇÕES COM SOLOS. **Scientia agraria, ISSN 1983-2443, Vol. 9, Nº. 3, 2008, pags. 317-329**, v. 9, 24 jun. 2008.

SOARES, D. B.; NÓBREGA, R. S.; GALVÍNCIO, J. D. Indicadores climáticos de desertificação na bacia hidrográfica do Rio Pajeú, Pernambuco. **Revista Brasileira de Climatologia**, v. 22, p. 380, 2018.

SOLARI, A. et al. Escavaçoes arqueológicas no abrigo funerário Pedra do Cachorro, Buíque – PE. **Clio Arqueológica**, v. 31, n. 1, p. 105–135, 2016.

SOUZA, T. F. DE. **Pinturas rupestres e paisagem: um estudo de caso das representações zoomórficas do Vale do Catimbau – PE**. [s.l.] Dissertação (Mestrado em Arqueologia) Programa de Pós-Graduação em Arqueologia - Universidade Federal de Pernambuco, Recife, 2016.

SPAULDING, A. C. Statistical Techniques for the Discovery of Artifact Types. **American Antiquity**, v. 18, n. 4, p. 305–313, 1953.

STEWARD, J. **Theory of culture change**. Urbana: University of Illinois Press, 1955.

TAVARES, B. **Evolução Morfotectônica dos Pedimentos Embutidos no Planalto da Borborema**. [s.l.] Tese (Doutorado em Geografia) Programa de Pós-Graduação em Geografia - Universiade Federal de Pernambuco, Recife, 2015.

THE-PANDAS-DEVELOPMENT-TEAM. **pandas-dev/pandas: Pandas 1.3.2**. Disponível em: <https://doi.org/10.5281/zenodo.3509134>.

TORRES, R. Arqueologia histórica na era digital. **Vestígios**, v. 11, n. 1, p. 7–19, 2017.

TRONCOSO, A.; MOYA, F.; BASILE, M. Rock art and social networks among hunter gatherers of north-central Chile. **Journal of Anthropological Archaeology**, v. 42, p. 154–

168, 2016.

USGS - UNITED STATES GEOLOGICAL SURVEY. **Earth Explorer - SRTM Data**. Disponível em: <https://earthexplorer.usgs.gov/>.

UTIDA, G. et al. Tropical South Atlantic influence on Northeastern Brazil precipitation and ITCZ displacement during the past 2300 years. **Scientific Reports**, v. 9, 8 fev. 2019.

VALLI, A. M. F.; MÜTZENBERG, D. **Reconstituição paleoambiental recente (Holoceno) da Lagoa Uri de Cima (Salgueiro, PE): uma contribuição da paleontologia.** (F. de L. R. Listo, D. Mützenberg, B. de A. C. Tavares, Eds.)E-book do I workshop de geomorfologia e geoarqueologia do Nordeste. **Anais**...Recife: Grupo de Estudos do Quaternário do Nordeste Brasileiro – GEQUA, 2016

VALLVERDU, J.; CASACUBERTA, D. Computational Philosophy as Experimental Philosophy. 13 out. 2019.

VAN HAVRE, G. **Interações: Análise da Complexidade no Registro Rupestre do Vale do Ventura, Morro do Chapéu, Bahia**. [s.l.] Tese (Doutorado em Arqueologia) Programa de Pós graduação em Arqueologia, Universidade Federal de Pewrnambuco. Recife-PE, 2015.

VAN RIJMENAMM, M. **The History Of Predictive Analytics - Infographic**. Disponível em: <https://datafloq.com/read/history-predictive-analytics-infographic/438>. Acesso em: 20 jul. 2020.

VERHAGEN, P. **Predictive Modelling**. (Axel Posluschny, K. Lambers, I. Herzog, Eds.)Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA). **Anais**...Berlin: 2008

VERHAGEN, P. et al. **Introducing the Human Factor in Predictive Modelling:: a Work in Progress**. (I. Romanowska et al., Eds.)CAA2012 Proceedings of the 40th Conference in Computer Applications and Quantitative Methods in Archaeology, Southampton, United Kingdom, 26-30 March 2012. **Anais**...Southampton: jan. 2012

VERHAGEN, P. Predictive Modelling. In: SANDRA L. LÓPEZ VARELA (Ed.). . **The Encyclopedia of Archaeological Sciences**. Newark: John Wiley and Sons, 2018. p. 1–3.

VERHAGEN, P.; BORSBOOM, A. The design of effective and efficient trial trenching

strategies for discovering archaeological sites. **Journal of Archaeological Science**, v. 36, n. 8, p. 1807–1815, 2009.

VERHAGEN, P.; WHITLEY, T. G. Integrating Archaeological Theory and Predictive Modeling: a Live Report from the Scene. **Journal of Archaeological Method and Theory**, v. 19, n. 1, p. 49–100, 2011.

VERHAGEN, P.; WHITLEY, T. G. Predictive Spatial Modelling. In: GILLINGS, M., HACIGÜZELLER, P. & LOCK, G. (Ed.). . **Archaeological Spatial Analysis: A Methodological Guide**. Nova Iorque: Routledge, 2020a. p. 231–246.

VERHAGEN, P.; WHITLEY, T. G. Predictive Spatial Modelling. In: GILLINGS, M.; HACIGÜZELLER, P.; LOCK, G. (Eds.). . **Archaeological Spatial Analysis: A Methodological Guide.** [s.l.] Taylor & Francis, 2020b. p. 231–246.

VITA-FINZI, C. et al. **Prehistoric Economy in the Mount Carmel Area of Palestine: Site Catchment Analysis**. (J. Gardiner, Ed.)Proceeding of the Prehistoric Society. **Anais**...Cambridge University Press, 1970

WACHTEL, I. et al. Predictive modeling for archaeological site locations: Comparing logistic regression and maximal entropy in north Israel and north-east China. **Journal of Archaeological Science**, v. 92, p. 28–36, 2018.

WEISS, A. D. . **Topographic Positions and Landforms Analysis**. ESRI International User Conference. **Anais**...San Diego: ESRI, 2001

WHEATLEY, D.; GILLINGS, M. **Spatial technology and archaeology: the archaeological applications of GIS**. [s.l: s.n.].

WHITE, D. A.; BARBER, S. B. Geospatial modeling of pedestrian transportation networks: a case study from precolumbian Oaxaca, Mexico. **Journal of Archaeological Science**, v. 39, n. 8, p. 2684–2696, 2012.

WILLEY, G.; PHILIP, P. **Method and Theory in American Archaeology**. Chicago: The University of Chicago Press, 1958.

WILLEY, G. R. **Prehistoric settlement patterns in the Viru Valley, Peru**. [s.l.] Bureau of American Ethnology Bulletin, 1953.

ZHU, X.; CHEN, F.; GUO, H. A Spatial Pattern Analysis of Frontier Passes in China's Northern Silk Road Region Using a Scale Optimization BLR Archaeological Predictive Model. **Heritage**, v. 1, p. 2, 20 mar. 2018.

ZUBROW, E. Digital Archaeology: a historical context. In: EVANS, T. L.; DALY, P. T. (Eds.). . **Digital Archaeology. Bridging Method and Theory**. Londres: [s.n.]. p. 10–31.

APÊNDICE A - CÓDIGOS E DADOS

Os dados e os códigos referentes a esta pesquisa podem ser encontrados nos endereços a seguir:

- códigos: https://github.com/lucasbonald/Pajeu_APM
- dados: DOI: 10.17632/v7gp2bmdfj.1.

APÊNDICE B – HIPERPARÂMETROS DO MODELO

Tabela 16: Hiperparâmetros usados para os cálculos dos modelos.

	HIPERPARÂMETROS
REGRESSÃO LOGÍSTICA	Penalidade = l2; iterações: 500; *solver* = saga;
MLP	ativação: RELU; alfa: 0,0001; tamanho das camadas escondidas = (50, 100, 50); taxa de aprendizado: constante; *solver* = adam
FLORESTA ALEATÓRIA	*bootstrap*= Falso; máxima profundidade= None; Número máximo de variáveis= auto; número mínimo de folhas = 1; número mínimo de divisões da amostra = 2, número de estimadores = 1000
FAB	*bootstrap*= Falso; máxima profundidade= None; Número máximo de variáveis= raiz quadrada; número mínimo de folhas = 1; número mínimo de divisões da amostra = 2, número de estimadores = 100; estratégia de reamostragem = maioria
BAGGING BALANCEADO	estratégia de reamostragem = maioria

Fonte: Elaborado pelo próprio autor (2022)

APÊNDICE C – VARIÁVEIS DO MODELO

Tabela 17: Quadro das variáveis usadas no modelo.

Variáveis Topográficas				
Inclinação	Orientação das Encostas	Índice de Posição Topográfica (IPT)	Perfil de Curvatura	Elevação
_slopemedian	*_aspectmedian*	*_tpimedian*	*_perfil_curvmedian*	*_demmedian*
_slopestdev	*_aspectstdev*	*_tpistdev*	*_perfil_curvstdev*	*_desstdev*
_slopemax	*_aspectmax*	*_tpimax*	*_perfil_curvmax*	*_demmax*
_slopemajority	*_aspectmajority*	*_tpimajority*	*_perfil_curvmajority*	*_demmajority*
_slopevariance	*_aspectvariance*	*_tpivariance*	*_perfil_curvvariance*	*_devvariance*

Variáveis Topográficas (Insolação)				
Difusa/Solstício de Verão	Difusa/Solstício de Inverno	Direta/Solstício de Verão	Direta/Solstício de Inverno	Total/Solstício de Verão
_diffuse_insol_svmedian	*_diffuse_insol_simedian*	*_direct_insol_svmedian*	*_direct_insol_simedian*	*_total_insol_svmedian*
_diffuse_insol_svstdev	*_diffuse_insol_sistdev*	*_direct_insol_svstdev*	*_direct_insol_sistdev*	*_total_insol_svstdev*
_diffuse_insol_svmax	*_diffuse_insol_simax*	*_direct_insol_svmax*	*_direct_insol_simax*	*_total_insol_svmax*
_diffuse_insol_svmajority	*_diffuse_insol_simajority*	*_direct_insol_svmajority*	*_direct_insol_simajority*	*_total_insol_svmajority*
_diffuse_insol_svvariance	*_diffuse_insol_sivariance*	*_direct_insol_svvariance*	*_direct_insol_sivariance*	*_total_insol_svvariance*
Total/Solstício de Inverno	Pôr do sol/Solstício de Verão	Pôr do sol/Solstício de Inverno	Nasc. do sol/Solstício de Verão	Nasc. do sol/Solstício de Inverno
_total_insol_simedian	*_sunset_svmedian*	*_sunset_simedian*	*_sunrise_svmedian*	*_sunrise_simedian*

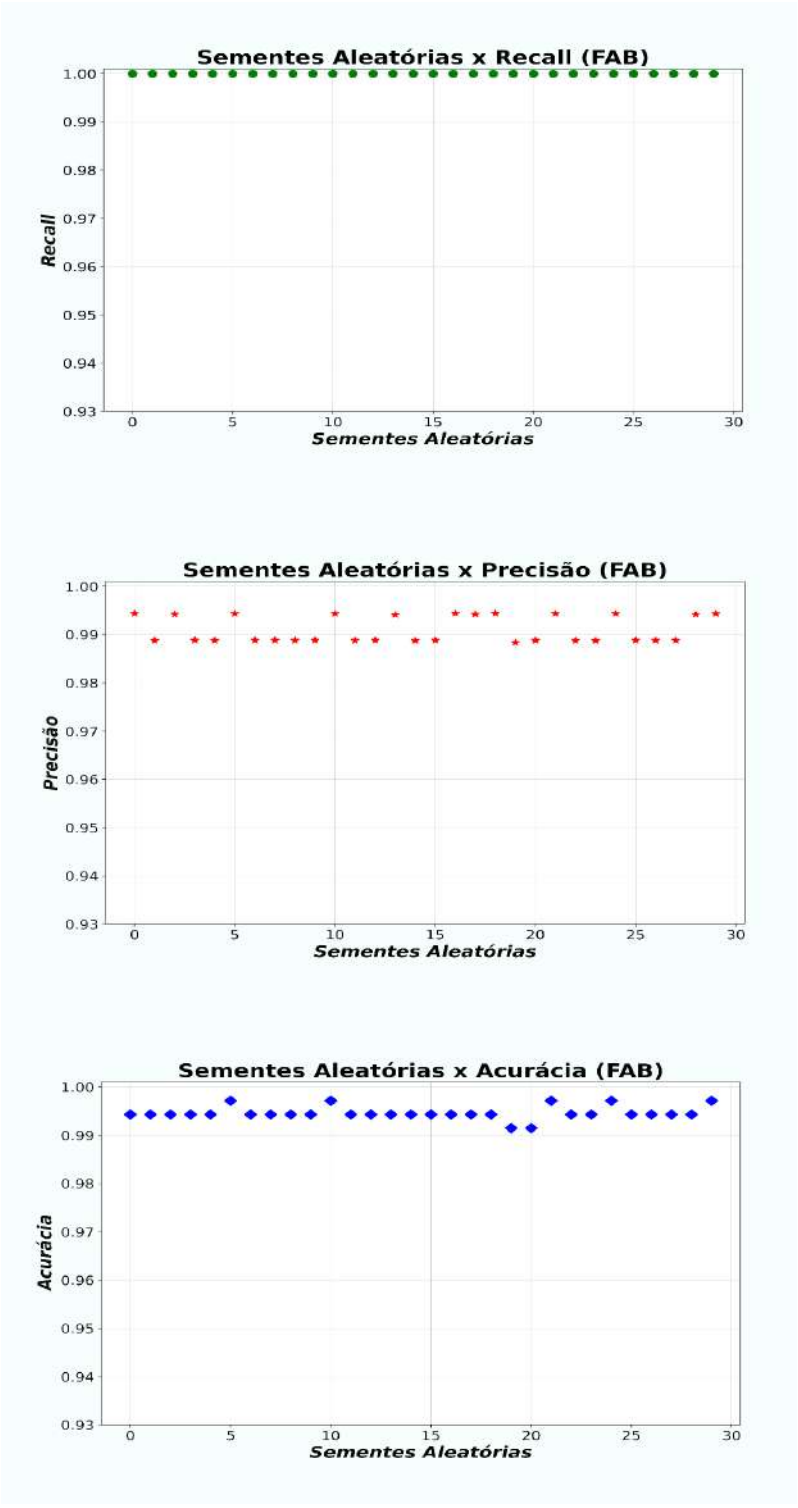
_total_insol_sistdev	*_sunset_svstdev*	*_sunset_sistdev*	*_sunrise_svstdev*	*_sunrise_sistdev*
_total_insol_simax	*_sunset_svmax*	*_sunset_simax*	*_sunrise_svmax*	*_sunrise_simax*
_total_insol_simajority	*_sunset_svmajority*	*_sunset_simajority*	*_sunrise_svmajority*	*_sunrise_simajority*
_total_insol_sivariance	*_sunset_svvariance*	*_sunset_sivariance*	*_sunrise_svvariance*	*_sunrise_sivariance*

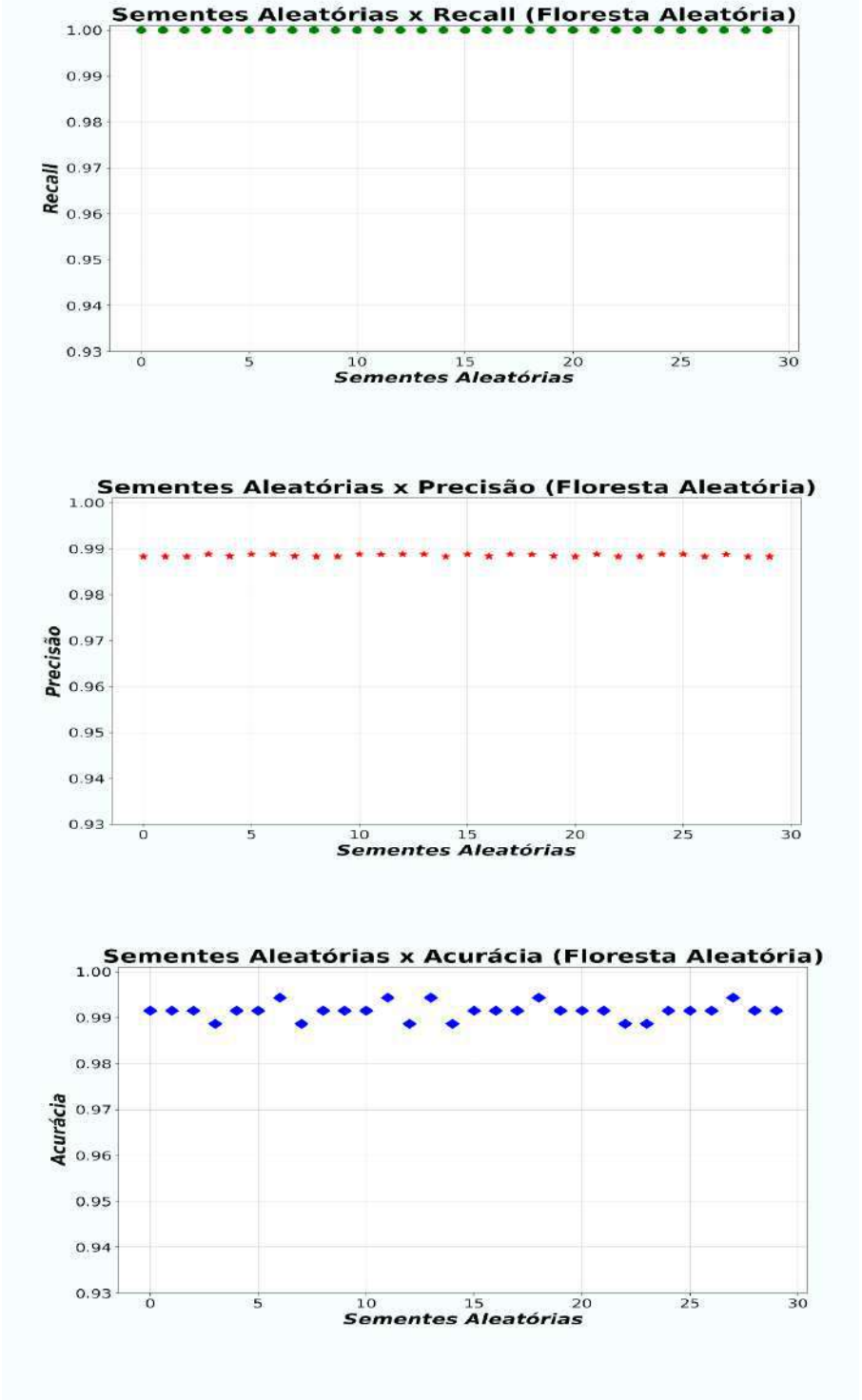
Variáveis sociais				
Índice de Visibilidade	Densidade/CMC insolação direta	Densidade/CMC Inclinação	Proximidade/CMC Insolação direta	Proximidade/CMC Inclinação
_vimedian	*_dens_insolmedian*	_dens_slopemedian	_prox_insolmedian	_prox_slopemedian
_vistdev	*_dens_insolstdev*	_dens_slopestdev	_prox_insolstdev	_prox_slopestdev
_vimax	*_dens_insolmax*	_dens_slopemax	_prox_insolmax	_prox_slopemax
_vimajority	*_dens_insolmajority*	_dens_slopemajority	_prox_insolmajority	_prox_slopemajority
_vivariance	*_dens_insolvariance*	_dens_slopevariance	_prox_insolvariance	_prox_slopevariance

Variáveis de captação de recursos			
Distância/principais drenagens	Distância/Rochas ígneas	Distância/Rochas Metamórficas	Distância/Rochas Sedimentares
_ordem_drenmedian	*_igneamedian*	*_metamedian*	*_sedmedian*
_ordem_drenstdev	*_igneastdev*	*_metastdev*	*_sedstdev*
_ordem_drenmax	*_igneamax*	*_metamax*	*_sedmax*
_ordem_drenmajority	*_igneamajority*	*_metamajority*	*_sedmajority*
_ordem_drenvariance	*_igneavariance*	*_metavariance*	*_sedvariance*

Fonte: Elaborado pelo próprio autor (2022)

APÊNDICE D – GRÁFICOS DE SEMENTES DO MODELO




Sementes Aleatórias x Recall (Floresta Aleatória)
Recall
Sementes Aleatórias
Sementes Aleatórias x Precisão (Floresta Aleatória)
Precisão
Sementes Aleatórias
Sementes Aleatórias x Acurácia (Floresta Aleatória)
Acurácia
Sementes Aleatórias