Predição da cobertura do solo baseada em um modelo de classificação*

Francine Moreira Ferreira[1 [0000−0002−8778−9299]],
Aline Pons Alves Lisboa[1 [0000−0001−5729−8205]], and
Sandro da Silva Camargo[1 [0000−0001−8871−3950]]

Universidade Federal do Pampa, Bagé RS, Brasil
{francineferreira.aluno, alinelisboa.aluno,
sandrocamargo}@unipampa.edu.br
https://cursos.unipampa.edu.br/cursos/ppgcap/

Resumo. O solo é um dos mais importantes recursos naturais do planeta. Um solo saudável é crucial para a manutenção da vida e bem estar dos seres vivos. A compreensão dos mapas de cobertura do solo é um elemento crítico na tomada de decisões e de gestão deste importante recurso natural. Para contribuir neste contexto, o presente trabalho apresenta um modelo de predição do tipo de cobertura do solo, com base em 148 atributos quantificados com base em imagens de alta resolução. O modelo apresentado atingiu uma acurácia de 68,05% nos dados de teste, sendo capaz de prever imagens referentes a asfalto, sombras, carros, edifícios, concreto, solo, piscinas e grama. Desta forma, o modelo proposto caracteriza-se como um recurso viável para análise automática de grandes volumes de imagens.

Palavras-Chave: Mineração de Dados · Modelagem · Árvores de Decisão · Uso da Terra.

1 Introdução

Compreender os mapas de cobertura do solo é um elemento crítico na tomada de decisões políticas, de desenvolvimento, de planejamento e de gestão de recursos. No entanto, analisar imagens de satélite de forma manual, ou baseada em esforço humano, é impraticável. Neste sentido se torna necessária a aplicação de recursos computacionais, baseados em inteligência artificial, que possam identificar autonomamente, sem trabalho humano, os elementos que estão nas imagens [1].

Para a tomada de decisão baseada em inteligência artificial, é estudada a área de Mineração de Dados (MD), que engloba tecnologias de bancos de dados, inteligência artificial, aprendizado de máquina, reconhecimento de padrões, entre outras funcionalidades, se tratanto, portanto, de uma área de pesquisa multidisciplinar que provêm da matemática, da estatística e da computação [2]. Com a

* Este trabalho foi desenvolvido com apoio financeiro da CAPES/FAPERGS (PDPG), Edital nº 197/2021.

incumbência do segmento analítico (*data analytics*) do *Big Data* e, ainda, pela composição de um amplo volume de dados e a análise propriamente dita - sendo percebida como componente de um processo mais abrangente - a exploração de conhecimentos em bases de dados, o que torna um dado o *input* para a formação do conhecimento [3].

Dentre as diversas tarefas de MD, se situa a Classificação, técnica utilizada no presente trabalho, que consiste na predição de uma variável categórica por meio de um modelo capaz de mapear um grupo de registros com relação a determinados atributos [4].

Assim, este trabalho tem como objetivo elaborar um modelo capaz de predizer qual o tipo de cobertura de solo baseado em atributos como área, índice de borda, brilho, comprimento, largura, índice de vegetação, assimetria, dentre outros, captados por imagens de alta resolução e quantificados na base de dados de estudo.

Se percebe a relevância da MD tanto neste contexto (análise de cobertura de solo) quanto em diversos outros, uma vez que, os grandes volumes de dados são evidenciados em diversos setores e negócios e representam, se corretamente analisados, "ouro" para as estratégias organizacionais. Além disso, muitas vezes, os usuários finais não são estatísticos, então, obtendo um modelo que já forneça dados categorizados e tratados, certamente irá facilitar conclusões e, consequentemente, tomadas de decisões mais assertivas [2].

Desta forma, percebendo a importância e o contexto de realização desse estudo como um todo, na sequência serão detalhados os aspectos referentes à sua realização. Na Seção 2 são tratados o material e os métodos utilizados para a concepção desse estudo, detalhando a base de dados analisada, as ferramentas e o as árvore de decisão. Logo após, na Seção 3, são demonstrados os resultados alcançados por meio da ferramenta e a respectiva análise estatística. Por fim, na Seção 4 são proferidas as conclusões atinentes.

2 Material e Métodos

Para a realização do presente estudo foi escolhida uma base de dados que considerasse características de imagens de alta resolução e que possibilitasse a análise da mesma utilizando a tarefa de Classificação. Para isso, foram aplicadas técnicas de MD com o intermédio do Software RStudio direcionadas a esta tarefa. A seguir são detalhados os aspectos referentes a cada um destes.

2.1 Base de Dados

O material de estudo do presente trabalho consiste em uma base de dados de cobertura do solo[8], a qual é disponibilizada para download no repositório digital *UCI Machine Learning - Center for Machine Learning and Intelligent Systems* [1]. Este banco de dados é direcionado à classificação da cobertura do solo por meio

[1] https://archive.ics.uci.edu/ml/datasets/Urban+Land+Cover

de imagens aéreas de alta resolução quanto aos diferentes atributos coletados nesta - possuindo como objetivo auxiliar nos esforços de um planejamento urbano sustentável - e foi elaborado por Brian Johnson, do Instituto de Estratégias Ambientais Globais no Japão [6, 7].

Quanto ao banco de dados, é composto por 147 atributos numéricos, e um atributo categórico, com a classe. Dentre os atributos numéricos, estão variáveis espectrais, de tamanho, de forma e de textura, repetidas em escalas. As classes das imagens aéreas são 9 tipos de cobertura de solo: árvores, grama, solo, concreto, asfalto, construções, carros, piscinas e sombras. Além disso, há 675 amostras divididas em dados de treinamento, com 168 amostras, e dados de teste, com 507 amostras. A quantidade de amostras por classe é apresentada na Tabela 1.

Table 1. Distribuição das amostras por classe nas bases de treino e teste.

Classe	Treino	Teste
Asfalto	14	45
Construções	25	97
Carros	15	21
Concreto	23	93
Grama	29	83
Piscina	15	14
Sombras	16	45
Solo	14	20
Árvore	17	89
Total	168	507

A Tabela 2 apresenta os atributos analisados neste estudo classificados conforme o que representam: variáveis de forma, de tamanho, espectrais e de textura. Quanto às variáveis de forma se tem: BrdIndx (Índice de Borda), Round (Redondeza), Compact (Compacidade), ShpIndx (Índice de Forma), LW (Comprimento/Largura), Rect (Retangularidade), Dens (Densidade), Assym (Assimetria) e BordLngth (Comprimento de Borda). Em relação ao tamanho, a variável área é a única observada nesta categoria. As variáveis espectrais são: Bright (Brilho), Mean_G (Média de Verde), Mean_R (Média de Vermelho), Mean_NIR (Desvio Padrão de Infravermelho Próximo) e NDVI (Índice de Vegetação de Diferença Normalizada). E as variáveis de textura: SD_G (Desvio Padrão de Verde), SD_R (Desvio Padrão de Vermelho), SD_NIR (Desvio Padrão de Infravermelho Próximo), GLCM1 (Matriz de Cooconcorrência de Nível de Cinza - MCNC), GLCM2 (MCNC - atributo 2) e GLCM3 (MCNC atributo 3). Estas variáveis possuem repetições em escala (40, 60, 80, 100, 120 e 140).

A seguir são especificadas as ferramentas utilizadas para a realização da classificação bem como o produto gerado para a análise dos dados.

Table 2. Classificação dos atributos da base de dados

Forma	Tamanho	Espectral	Textura
BrdIndx	Área	Bright	SD_G
Round		Mean_G	SD_R
Compact		Mean_R	SD_NIR
ShpIndx		Mean_NIR	GLCM1
LW		NDVI	GLCM2
Rect			GLCM3
Dens			
Assym			
BordLngth			

2.2 Ferramentas

Para a realização do trabalho, foram utilizados os recursos do ambiente *RStudio* versão 4.1.1 sobre o sistema operacional Windows em plataforma x86 64 bits. Foram também utilizados os pacotes *Classification and Regression Training* (Caret) versão 6.0-92 e *Rpart.plot* versão 3.1.1. O pacote *caret* foi utilizado pois é direcionado às técnicas de Classificação e Regressão, promovendo o particionamento de dados, de forma a propiciar resultados como a Matriz de Confusão para os modelos de treino e de teste, a qual aponta o número de observações corretamente preditas e também os equívocos e em relação a que outras categorias estes ocorreram. Já o pacote *Rpart.plot* foi utilizado com vistas à geração de uma Árvore de Decisão do modelo concebido.

2.3 Árvore de Decisão

As árvores de decisão representam uma estrutura constituída por ramos que representam os atributos contidos na base de dados analisada. Nesta estrutura, cada nó interno é um atributo distinto do atributo-classe que representam uma decisão por associação entre o atributo e a variável alvo. Além disso, as folhas simbolizam valores atribuídos ao atributo classe. Neste cenário, a finalidade desta é obter uma estrutura preditiva para o problema em questão [2, 5]. Em seguida são apresentados o modelo e os resultados alcançados.

3 Resultados e Discussão

Os resultados apresentados nessa sessão serão discutidos por meio da árvore de decisão gerada e também pela análise estatística.

3.1 Árvore de Decisão e Matriz de Confusão

A Figura 1 apresenta o modelo de árvores de decisão criada. Por meio da visualização desta se pode inferir que se o atributo NDVI não é $< 0{,}055$ e o atributo

Bright_80 é $\geq$ 137, o resultado tem 88% de probabilidade de se referir a uma imagem de grama (Figura 2). Por outro lado, se o NDVI não é $<$ 0,055 e o atributo Bright_80 não é $\geq$ 137, somados ao fato do atributo Rect ser $\geq$ 0,8, então 86% de probabilidade de ser uma sombra e, em caso negativo de Rect (Rect não ser $\geq 0,8$), então há 100% de probabilidade da imagem representar árvores (Figura 3).

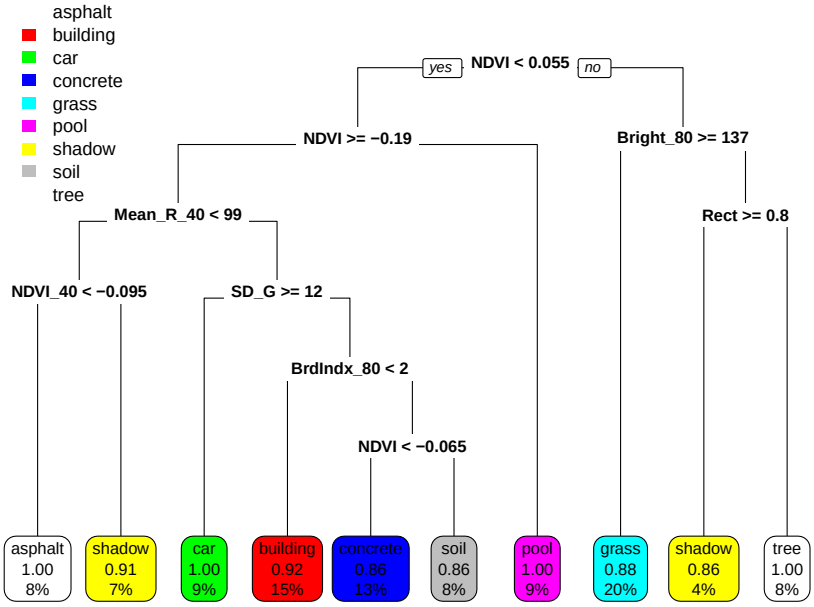

Fig. 1. Modelo preditivo de árvores de decisão.

Por outro lado, caso o NDVI seja $<$ 0,055 mas não seja $\geq$ -0,19, então há 100% de possibilidade da composição urbana ser relacionada a piscina (Figura 4). Nesta mesma linha, ainda considerando o NDVI $\geq$ -0,19 mas, agora, o agregando ao atributo Mean_R_40 $<$ 99 e considerando também que o NDVI_40 seja $<$ -0,095, se tem que em 100% dos casos, as imagens serão compatíveis com

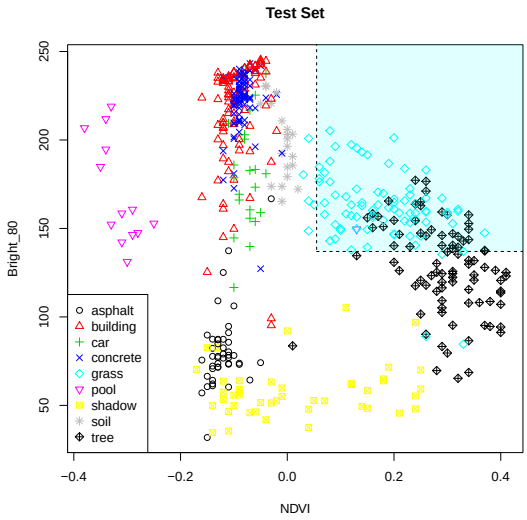

Fig. 2. Superfície de Decisão da regra da árvore de decisão que classifica **grama** com NDVI>=0.055 e Bright_80>=137.

asfalto. Já na condição do NDVI_40 não ser < -0,095, há 91% de probabilidade da imagenm representar uma sombra (Figura 5).

Em sequência, considerando que o atributo Mean_R_40 não é < 99 e SD_G é ≥ 12, assim 100% das imagens são relativas a carros. Ao passo que se considerarmos que Mean_R_40 não é < 99 aliado ao atributo SD_G não ser ≥ 12 e BrdIndx_80 < 2, destarte há 92% de probabilidade da imagem se referir a edifícios. Por outra perspectiva, se forem consideradas negativas para os nós Mean_R_40 < 99, SD_G ≥ 12, BrdIndx_80 < 2 e NDVI < -0,065, há 86% de probabilidade da imagem se referir a solo. E, de uma perspectiva similar, porém, considerando que o NDVI é < -0,065, há 86% de probabilidade da imagem representar concreto. Estas regras não foram representadas em gráficos específicos por envolverem a presença de vários atributos, impossibilitando sua apresentação em um plano bidimensional.

No modelo de treinamento, a Matriz de Confusão mostrou que das 14 imagens de asfalto, 13 o modelo foi capaz de prever e 1 confundiu com sombra. De 25 edifícios, 23 foram previstos com precisão e em apenas 2 houve imprecisão, confundindo com concreto. Dos 15 carros, todos foram previstos com sucesso. Das 23 imagens de concreto, 19 foram preditas com sucesso, 2 associadas a edifícios e 2 ao solo. Das 29 imagens de grama, todas foram previstas desse modo.

Por fim, das 15 imagens de piscina, todas foram detectadas corretamente. Das 16 imagens de sombra, todas foram corretamente associadas também. Quanto às imagens referentes ao solo apresentaram 12 resultados corretamente relacionados e dois associados à grama. E, das 17 imagens de árvores, 13 estavam perfeitamente associadas, 1 estava associada ao concreto, duas à grama e uma outra es-

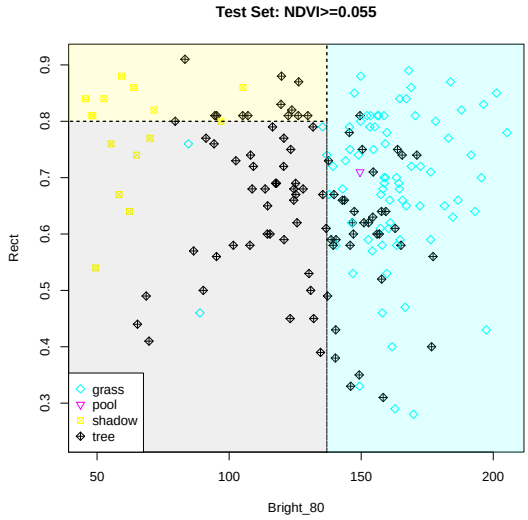

Fig. 3. Superfície de Decisão da regra da árvore de decisão que classifica **árvore** com NDVI>=0.055, Bright_80<137 e Rect<0.8. Regra da árvore de decisão que classifica **sombra** com NDVI>=0.055, Bright_80<137 e Rect>=0.8.

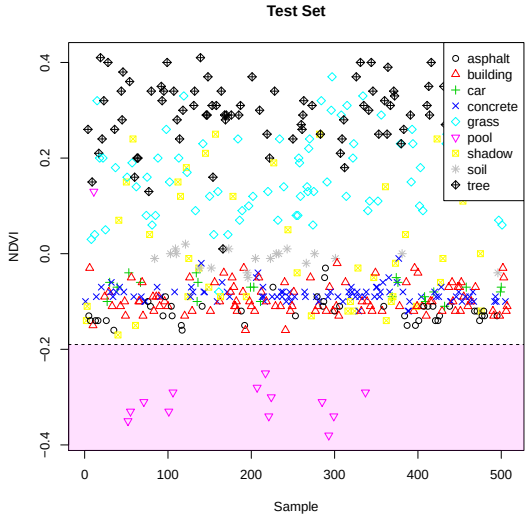

Fig. 4. Superfície de Decisão da regra da árvore de decisão que classifica **piscina** com NDVI<-0.19.

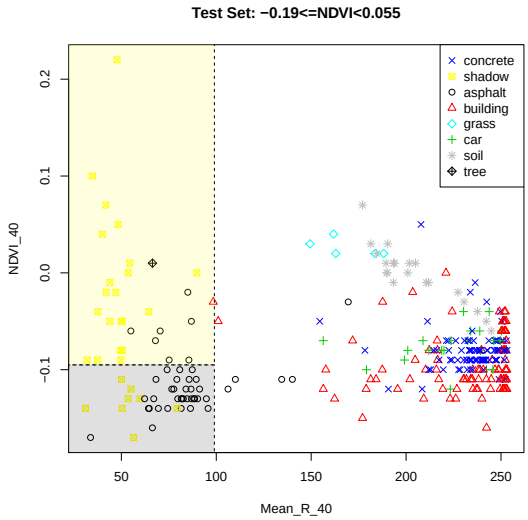

Fig. 5. Superfície de Decisão da regra da árvore de decisão que classifica **asfalto** com NDVI entre -0.019 e 0.055, Mean_R<99 e NDVI_40<-0.095. Regra da árvore de decisão que classifica **sombra** com NDVI entre -0.019 e 0.055, Mean_R<99 e NDVI_40>=-0.095.

tava relacionada à sombra. Este modelo (conjunto de treino) apresentou 92,26% de acurácia nas predições e Valor P $<$ 2.2e-16.

Já no modelo de teste, a Matriz de Confusão mostrou que das 45 imagens de asfalto, o modelo foi capaz de prever 33, associou uma a edifício, duas a carros e também duas a contreto; dos 97 edifícios, 56 deles assertivamente relacionados e 41 erros, associando 7 a carros, 25 a concreto, um a sombra e oito a solo. Dos 21 carros, 17 corretamente antevistos e 4 confundidos com concreto; das 93 imagens relativas a concreto, 70 foram visualizadas com sucesso, 13 foram associada a edifícios e 7 a carros; das 83 imagens de grama, 73 prognosticadas assim, duas geraram confusão com carro, uma com concreto, três com o solo e quatro com árvores.

Finalmente, das 14 imagens de piscinas, 13 foram detectadas corretamente e uma foi relacionada à grama. Das 45 imagens de sombra, 30 foram corretamente associadas, 8 classificadas como asfalto e sete como árvore. As imagens do solo apresentaram 13 resultados corretamente relacionados, quatro considerados edifícios e três julgados como árvores. Por fim, das 89 imagens de árvores, 40 foram perfeitamente compatíveis, 35 classificadas como grama e 14 com sombra. Este modelo (conjunto de teste) apresentou 68,05% de acurácia e valor P $<$ 2.2e-16.

Table 3. Matriz de Confusão do modelo aplicado sobre os dados de teste.

Classe	Classe Real								
Predita	Asfalto	Construção	Carro	Concreto	Grama	Piscina	Sombra	Solo	Árvore
Asfalto	**33**	0	0	0	0	0	8	0	0
Construção	1	**56**	0	13	0	0	0	4	0
Carro	2	7	**17**	7	2	0	0	3	0
Concreto	2	25	4	**70**	1	0	0	0	0
Grama	0	0	0	0	**73**	1	0	0	35
Piscina	0	0	0	0	0	**13**	0	0	0
Sombra	7	1	0	0	0	0	**30**	0	14
Solo	0	8	0	3	3	0	0	**13**	0
Árvore	0	0	0	0	4	0	7	0	**40**
Total	45	97	21	93	83	14	45	20	89

Table 4. Métricas do modelo aplicado sobre os dados de teste.

Métrica	Valor
Accuracy	0.6805
95% CI	(0.6379, 0.7209)
No Information Rate	0.1913
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.6285

3.2 Análise Estatística

A Tabela 5 apresenta a estatística descritiva dos sete atributos mais relevantes identificados pelo modelo deste estudo. Para cada atributo foram calculados: valor mínimo, primeiro quartil, mediana, média, terceiro quartil, valor máximo, normal e valor F.

Table 5. Estatística Descritiva do Modelo de Composição Urbana

Atributo	Mínimo	1 Quartil	Mediana	Média	3 Quartil	Máximo	Normal	Valor F
BrdIndx_80	1	1.57	2.455	2.5726	3.3625	5.28	S	14.4654
Bright_80	41.2	124.2925	161.515	159.8515	215.2075	244.85	S	76.321
Mean_R_40	33.34	100.275	157.51	161.7056	232.7975	252.71	S	92.128
SD_G	4.33	6.77	8.01	10.1314	11.5	36.4	S	27.5447
Rect	0.22	0.67	0.78	0.7476	0.84	1	S	6.0571
NDVI	-0.36	-0.1	-0.065	-0.0031	0.1	0.39	S	118.9482
NDVI_40	-0.34	-0.1	-0.05	0.0015	0.0925	0.39	S	118.2045

Neste recorte constam os atributos iniciais NDVI, Rect e SD_G e também a repetição na escala 40 para o atributo Mean_R e NDVI e na escala 80 para os atributos BrdIndx e Bright, retratando os atributos que, em conjunto, foram percebidos como os sete atributos mais relevantes pelo algoritmo que originou

a árvore de decisão (Figura 1), conforme já detalhado anteriormente na representação da árvore de decisão.

4 Conclusões

O modelo foi capaz de diagnosticar a porcentagem de presença dos nove componentes urbanos analisados, em que puderam ser observados na seguinte ordem crescente de identificação: asfalto, árvores e solo (8% cada um), piscinas e carros (9% cada), sombras e concreto (13% cada), 15% de edifícios e 20% de grama.

Dos 148 atributos em estudo, os mais relevantes (que melhor explicam o problema) detectados pelo modelo foram: NDVI escala original e 40 (Índice de Vegetação por Diferença Normalizada), Bright_80 (Brilho), Mean_R_40 (Vermelho), SD_G (Desvio Padrão do Verde), Rect (Retangularidade) e BrdIndx_80 (Índice de fronteira). Destas, quatro são variáveis espectrais, uma variável é de textura e duas são de forma.

Deste modo, se considera que o trabalho conseguiu alcançar o seu objetivo, o qual visava predizer determinados tipos de cobertura de solo baseado nos 148 atributos analisados. Dentro deste contexto, se ressalta que a utilização de MD, mais especificamente a Classificação, auxiliou quanto aos esforços direcionados ao planejamento urbano sustentável nessa localidade, uma vez que forneceu uma percepção em termos de constituição relativa aos tipos de cobertura, o que pode contribuir com o planejamento desta cidade com associado ao que pretende ter em sua composição urbana de modo que se torne mais sustentável ao longo dos anos. Do mesmo modo pode vir a auxiliar em inúmeros outros contextos e estudos posteriores, como a detecção de pragas em lavouras e potenciais áreas de plantação, por exemplo.

References

1. Saah, D., et al.: Land Cover Mapping in Data Scarce Environments: Challenges and Opportunities. Frontiers in Environmental Science **7** (2019) Disponível em https://www.frontiersin.org/article/10.3389/fenvs.2019.00150. doi:10.3389/fenvs.2019.00150
2. De Amo, S.: Técnicas de Mineração de Dados. Jornada de Atualização em Informática 2004, p. 26.
3. Ferrari, D. G.; Silva, L. N. de C.: Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações. Saraiva Educação SA (2017)
4. Galvão, N. D.; Marin, H. de F.: Técnica de mineração de dados: uma revisão da literatura. Acta Paulista de Enfermagem, v. 22, p. 686-690, 2017.
5. Ochôa, F. Jimenez, et al.: Triagem automatizada de pacientes com risco de Câncer de Mama. In: Congreso Argentino de Informatica Y Salud (CAIS). pp. 49–58. Sadio, Buenos Aires, Argentina (2021)
6. Johnson, B.; Xie, Z.: Classifying a high resolution image of an urban area using super-object information. ISPRS Journal of Photogrammetry and Remote Sensing, 83, 40-49, 2013.

7. Johnson, B.: High resolution urban land cover classification using a competitive multi-scale object-based approach. Remote Sensing Letters, 4 (2), 131-140, 2013.
8. *UCI Machine Learning Repository - Urban Land Cover Data Set*, https://archive.ics.uci.edu/ml/datasets/Urban+Land+Cover. Last accessed 13 Jun 2022.