e-Acadêmica, v. 3, n. 3, e0433267, 2022
(CC BY 4.0) | ISSN 2675-8539 | DOI: http://dx.doi.org/10.52076/eacad-v3i3.267

Alternativas locacionais de empreendimentos utilizando aprendizado de máquina

Machine learning to location analysis

Recebido: 20/08/2022 | Revisado: 29/08/2022 | Aceito: 31/08/2022 | Publicado: 02/09/2022

Hebert Ramos
ORCID: https://orcid.org/0000-0002-3955-337X
Universidade de São Paulo, Brasil
E-mail: hebertramos@yhaoo.com.br
Vladimir Diniz
ORCID: https://orcid.org/0000-0002-8686-1658
Pontifícia Universidade Católica de Minas Gerais, Brasil
E-mail: vladimir.diniz@gmail.com

Resumo
A Ciência de Dados pode ser descrita como uma área multidisciplinar voltada para a estruturação e análise de dados, tendo como objetivo extrair conhecimento para a tomada de decisão. Decidir a melhor localização para um projeto ou empreendimento pode ser uma tarefa extremamente complexa, na medida em que precisa considerar diversas variáveis, muitas das quais com características conflitantes. Dentro da Ciência de Dados o Aprendizado de Máquina é amplamente utilizado para lidar com problemas de diversas áreas de aplicação, onde os aspectos espaciais são essenciais como a análise de alternativa locacional, análise de transporte, crescimento urbano, predição em agricultura, dentre outras. Porém as propriedades dos dados espaciais frequentemente são ignoradas pelos algoritmos de Aprendizado de Máquina, em domínios de aplicações espaciais. NIKPARVAR e THILL 2021, argumentam que os dados espaciais apresentam certas propriedades distintas que os diferenciam de outros tipos de dados, como dependência espacial, heterogeneidade espacial e escala. Quando é necessário realizar estudos que envolvem as variáveis e as características de determinado território, torna-se indispensável analisar os dados espaciais. Este trabalho avalia a aplicação de Aprendizado de Máquina como alternativa para determinar a melhor localização para um projeto ou empreendimento. O trabalho busca utilizar uma abordagem que possa ser empregada de forma complementar a métodos já consagrados de geoprocessamento e análise multicritério.
Palavras-chave: Aprendizado de máquina; Geoprocessamento; Alternativa locacional; Regressão; Redes neurais.

Abstract
Data Science can be understood as a multidisciplinary area to data analysis and to decision making support. Deciding the best location for a project or enterprise can be an extremely complex task, as it needs to consider several variables, many of which having conflicting characteristics. Within the Data Science, Machine Learning is used to deal with several application areas, where spatial aspects are essential, such as location analysis, urban growth an agriculture. However, the properties of spatial data are often underrated for Machines Learning algorithms. NIKPARVAR and THILL 2021 argue that spatial data exhibit distinct properties that differentiate them, such as spatial heterogeneity, spatial dependence, and scale. When it is necessary to carry out studies that involves understanding the territory, it becomes necessary to analyze the spatial data. This paper presents a Machine Learning approach as an alternative to determine the best location for a project or enterprise, that can be complementary to geoprocessing and multicriteria analysis.
Keywords: Machine learning; Geoprocessing; Location alternative; Regression; Neural networks.

1. Introdução

Já se passaram anos desde que as questões ambientais e de sustentabilidade entraram definitivamente na pauta da sociedade. Mais recentemente o termo ESG (Environmental, Social and Governance) alcançou o mainstream empresarial com empreendedores e tomadores de decisão cada vez menos disposto a investir em empresas pouco sustentáveis do ponto de vista ambiental, social e de governança.

Um aspecto extremamente importante quando se avaliam investimentos públicos e privados são os impactos ambientais e dentro desses, a escolha do local do empreendimento ou atividade. À essa avaliação dá-se o nome de análise de alternativa

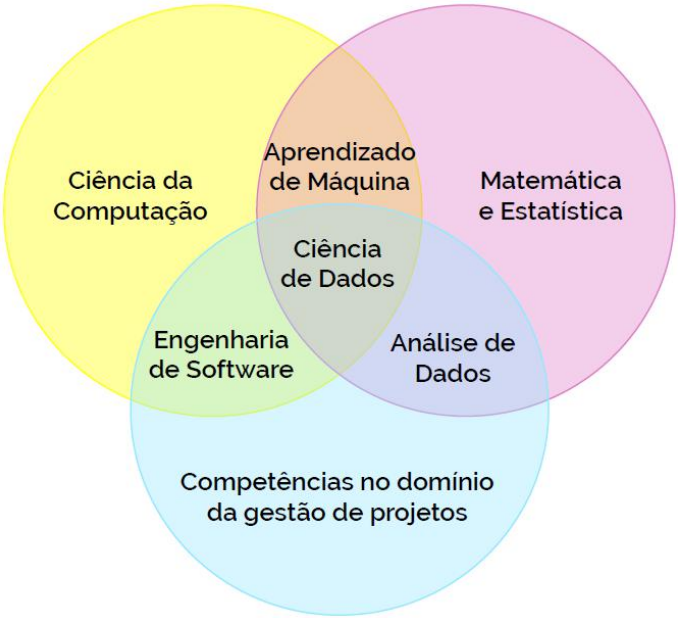
locacional.

Segundo Mattos (2018) apesar da análise de alternativa locacional ter um papel primordial nos possíveis impactos ambientais, normalmente é realizada de forma inadequada, incompleta e apresentando baixa qualidade.

De acordo com (Fernandes et al., 2017), na maioria dos casos, a análise se restringe à avaliação de uma única alternativa, não sendo realizado o confronto de opções. Isso, por sua vez, reduz a efetividade do processo tornando possível apenas a mitigação de impactos, que de outra forma poderiam ser evitados.

A ciência de dados, por sua vez, pode ser descrita como uma área interdisciplinar voltada para o estudo e a análise de dados econômicos, financeiros e sociais, estruturados e não-estruturados, como o objetivo de extrair conhecimento, detectar padrões e obter de insights para suporte à tomada de decisão. A Figura 1 apresenta um esquema conceitual da ciência de dados. É importante notar que a ciência de dados está na intercessão de diversas disciplinas, distribuídas entre ciência da computação, matemática e gestão de projetos.

Figura 1 - O que é ciência de dados.



Fonte: Apostila de Introdução à Ciência de Dados - CeMEAI-USP.

A utilização da ciência de dados tem grande potencial para dar suporte, gerar insights e contribuir de forma efetiva para que as empresas se tornem cada mais sustentáveis e competitivas.

Este trabalho tem como objetivo testar uma abordagem de ciência de dados que contribua para a sustentabilidade de empreendimentos. O capítulo 2 apresenta uma revisão bibliográfica, não extensiva, de assuntos pertinentes ao tema como Análise de Alternativa Locacional, Geoprocessamento e Aprendizado de máquina.

No capítulo 3 é proposta uma abordagem metodológica baseada em aprendizado supervisionado com algoritmos de regressão, métricas de performance de modelos, coleta, processamento e estruturação dos dados.

O capítulo 3 apresenta e discute os resultados obtidos pelo estudo e o capítulo 4 apresenta as principais conclusões do trabalho.

Este trabalho tem como objetivo testar uma abordagem de ciência de dados que contribua para a sustentabilidade de empreendimentos. O capítulo 2 apresenta uma revisão bibliográfica, não extensiva, de assuntos pertinentes ao tema como Análise de Alternativa Locacional, Geoprocessamento e Aprendizado de máquina.

No capítulo 3 é proposta uma abordagem metodológica baseada em aprendizado supervisionado com algoritmos de

(CC BY 4.0) | ISSN 2675-8539 | DOI: http://dx.doi.org/10.52076/eacad-v3i3.267

regressão, métricas de performance de modelos, coleta, processamento e estruturação dos dados.

O capítulo 3 apresenta e discute os resultados obtidos pelo estudo e o capítulo 4 apresenta as principais conclusões do trabalho.

2. Revisão Bibliográfica

A pressão para que as empresas estejam comprometidas com práticas sustentáveis sob o ponto vista ambiental, social e de governança tem aumentado tremendamente.

Relacionada às questões ambientais, a escolha do local mais adequado para a implantação de um projeto ou início de uma atividade torna-se crítica e deve considerar os aspectos socioambientais, além dos técnico-econômicos. O estudo das alternativas de localização é tratado por muitos como o "coração" do processo de Avaliação de Impacto Ambiental (AIA), através do qual o empreendedor busca demostrar aos interessados a viabilidade ambiental do seu empreendimento (Furlanetto, 2012).

A proposição de alternativas mais sustentáveis é considerada um dos princípios de boas práticas da AIA. Sem uma avalição efetiva de alternativas locacionais, a AIA pode se reduzir à proposição de ações de mitigação de impactos que poderiam ser evitados (Fernandes et al., 2017).

A AIA é um instrumento utilizado mundialmente para avalição dos impactos da ação humana sobre o meio. Criada nos Estados Unidos no final da década de 60, a AIA descreve processos para identificação, avaliação, prevenção, e mitigação dos efeitos de empreendimentos que sejam potencialmente danosos para o meio ambiente e para a sociedade, antes de sua ocorrência (Iaia, 1999 apud Fernandes et al., 2017).

A adoção da AIA no Brasil se iniciou em 1981 através da Lei Federal nº 6.938, Política Nacional do Meio Ambiente (PNMA). A PNMA tem como principal objetivo promover a preservação, a melhoria e a recuperação da qualidade ambiental, assegurando as condições necessárias ao desenvolvimento socioeconômico, aos interesses da segurança nacional e à proteção da dignidade da vida humana. Dentre os vários instrumentos implementados pela PNMA inclui-se o processo de avaliação de impactos e o estudo de alternativas locacionais (Mattos, 2018).

Apesar da reconhecida importância e da ampla utilização da AIA, diversos autores apontam falhas e inconsistências na maioria dos processos de estudos de alternativa locacional (Mattos, 2018). A Figura 2 apresenta uma relação das principais deficiências e os autores que as apontam, todas as referências são enfáticas ao destacar a baixa qualidade dos estudos, seja pela quantidade insuficiente, ou pela inadequação do conteúdo.

(CC BY 4.0) | ISSN 2675-8539 | DOI: http://dx.doi.org/10.52076/eacad-v3i3.267

Figura 2 - Deficiências dos processos de análise de alternativa locacional.

Deficiências	Referências
Não consideração do cenário de não execução	Caldas, 2006; Smith, 2007; Fernandes et. al., 2017; Schoen et al., 2016.
Ausência ou número insuficiente de alternativas	BRASIL, 2004; Smith, 2007; Pinho, Maia e Monterroso, 2007; Momtaz e Kabir, 2013; Almeida e Montaño, 2017
Alternativas impraticáveis, inconsistentes ou reconhecidamente inferiores a selecionada	Clark e Canter, 1997; Zubair, 2001; BRASIL, 2004; Smith, 2007; Machado, 2015; Fernandes et. al., 2017; Enríquez-de-Salamanca, 2018.
Baixa qualidade dos estudos de alternativas	Glasson et al., 1997; Gray e Edwards-Jones, 2003; Canelas et al., 2005; Pinho, Maia e Monterroso, 2007; Caldas, 2006; Momtaz e Kabir, 2013
Análise ou comparação incompletas das alternativas	Lee e Colley, 1992; Lee et al. 1999; Clark e Canter, 1997; Gray e Edwards-Jones, 2003; BRASIL, 2004; Smith, 2007; Meireles, 2011; Almeida et al., 2012; Landim e Sánchez, 2012; Pinho, Maia e Monterroso, 2007; Momtaz e Kabir, 2013; Hapuarachchi, Hughey, Rennie, 2016; Almeida e Montaño, 2017
Momento de decisão da localização anterior a realização da AIA	Steinemann, 2001; Benson, 2003; Momtaz e Kabir, 2013; Sánchez, 2013; Naser, 2015; Okubo, 2016, Khosravi, Jha-Thakur, Fischer, 2019.
Ausência ou fraca justificativa da escolha locacional	BRASIL, 2004; Gray e Edwards-Jones, 2003; Pinho, Maia e Monterroso, 2007; Almeida e Montaño, 2017; Fernandes et al., 2017.

Fonte: Mattos (2018).

Para além das questões socioambientais a localização geográfica é um recurso chave para projetos, para atividades e para empreendimentos. Quando somadas aos demais recursos como recursos financeiros, de capital humano, organizacionais e tecnológicos podem diferenciar uma empresa em relação aos seus concorrentes. Em contrapartida quando a localização geográfica é negligenciada como um recurso, pode tornar-se um dificultador para o crescimento ou mesmo inviabilizar a execução de determinadas atividades. Dessa forma, determinar a localização de um projeto ou empreendimento requer avaliações estratégicas, realização de estudos criteriosos, e não apenas a utilização de parâmetros subjetivos (Carnasciali & Delazari, 2011).

A determinação da melhor localização para um projeto ou empreendimento envolve grande número de variáveis, critérios de avaliação complexos e muitas vezes conflitantes. Neste contexto as ferramentas de Geoprocessamento, juntamente com as técnicas de análise multicritério (MCDM – Multicriteria Decision Making) vem sendo usadas como poderosas ferramentas para lidar com os problemas de localização (Zambon, 2004).

A definição do termo Geoprocessamento pode ser bastante abrangente, mas no geral se refere a todo um conjunto de técnicas utilizadas para lidar com informações geográficas. Dentro desse contexto os Sistemas de Informação Geográfica ou SIGs, são sistemas de informação com recursos especiais para manipular informação georreferenciada (Davis & Fonseca, 2001).

Nos SIGs a representação dos objetos se divide em duas grandes classes; geo-campos utilizados para representar

fenômenos de variação contínua e geo-objetos, utilizados para representar entidades individualizáveis. A representação destes fenômenos em nível lógico e físico é feita através de estruturas denominadas vetores ou imagens.

No modelo vetorial a localização e a representação dos objetos e feita por pares de coordenadas, pelas estruturas básicas ponto, linha e polígono e por combinações destas.

O ponto é um par ordenado (x, y) de coordenadas espaciais, sendo a forma mais simples de representação dos geo-campos permitindo apenas a localização dos mesmos.

De forma resumida a linha pode ser entendida como um segmento que uni dois pontos. A linha é utilizada para representar objetos cujo comprimento é muito maior que a largura como estradas, rios e ferrovias.

O polígono é uma região do plano delimitada por uma linha poligonal, sendo usados para representar objetos individualizáveis cujas áreas são relevantes, como edificações, propriedades, áreas de preservação.

A combinação das estruturas ponto, linha e polígono permite criar estruturas mais complexas como redes, isolinhas, tesselação, dentre outras. Para este trabalho optou-se por uma abordagem simplificada de definição das estruturas de representação utilizadas pelos SIGs, para uma visão detalhada consultar (Davis &Fonseca, 2001).

Outra forma de representação adotada pelos SIGs é a matricial, através de imagens. Uma imagem digital pode ser definida como uma função (x,y), bidimensional, válida em uma região. Essa função assume apenas valores positivos inteiros. A região em que a função é definida constitui uma matriz regular de pontos, daí o termo estrutura matricial. O processo de acessar uma sequência de valores em uma matriz geralmente se dá na forma de uma varredura (por exemplo, da esquerda para a direta e de cima para baixo), daí a utilização do termo raster (varredura em inglês) (Davis & Fonseca, 2001).

O formato de representação matricial ou raster permite a aplicação de uma série de algoritmos e técnicas matemáticas sobre os dados.

A combinação de diferentes técnicas de análise multicritério com os SIGs tem sido amplamente utilizada para tratar problemas de localização na medida em que permite conhecer e sumarizar relações espaciais e dessa forma explorar esse conhecimento para estruturar soluções e modelos de localização (Murray, 2010, apud Spiglon, 2015).

Como por exemplo o trabalho de Souza et al. (2020), cujo objetivo foi avaliar a distribuição espacial dos portos secos em Minas Gerais e identificar os melhores locais para uma nova instalação. A avaliação foi subsidiada através dos métodos de Análise Multicritério Espacial e Problema de Localização-Alocação, sendo que esse segundo se trata de uma aplicação do problema de p-medianas, sendo amplamente empregado na localização de instalações logísticas (Lorena et al., 2001 apud Souza et al., 2020).

A análise para a definição dos locais para instalação dos portos secos considerou critério ambientais, de competição, econômicos, logísticos e sociais. Os dados e suas respectivas fontes estão listados na Figura 3, ressalta-se que todos dados foram obtidos de fontes oficiais. Informações sobre metodologia utilizada e sobre os resultados da análise podem ser e encontrados em (Souza et al., 2020).

Figura 3 - Alternativa locacional para portos secos: Critérios selecionados e fonte de dados.

Grupo	*Variável*	*Fonte*	*Data*	*Formato dado*	*Método*
Econômico	Valor FOB exp	MDIC	2015	Tabular	Interpolação (IDW)
	Valor FOB imp	MDIC	2015	Tabular	Interpolação (IDW)
	nº postos de trabalho	RAIS	2010	Tabular	Interpolação (IDW)
	PIB	FGV	2015	Tabular	Interpolação (IDW)
Competição	Distância PS existente	PNLT 2010	2010	Vetor - Ponto	Distância Euclidiana
Acessibilidade	Dist. Rod. Troncal	DEER/Setop	2015	Vetor - Linha	Dist. Euclidiana
	Dist. Rod. Principal	DEER/Setop	2015	Vetor - Linha	Dist. Euclidiana
	Distância ferrovias	PNLT 2010	2010	Vetor - Linha	Dist. Euclidiana
Social	nº de não ocupados	IBGE	2010	Tabular	Interpolação (IDW)
	Índice Social	Atlas Brasil	2010	Tabular	Interpolação (IDW)
Ambiental	Dist. área urbana*	Embrapa Territorial	2015	Vector - Polígono	Dist. Euclidiana
	Dist. Hidrografia**	IBGE	2015	Vector - Linha	Dist. Euclidiana
	Dist. Ucs***	ICMBio/Semad	2015	Vector - Polígono	Dist. Euclidiana

*Influência limitada a 5 km da área urbanizada.

**Apesar da Lei Nº 12.651, de 25 de maio de 2012, que instituiu o novo Código Florestal (BRASIL, 2012) determinar que a maior Área de Proteção Permanente (APP) de mata ciliar é de 500 m, foi definida uma oneração gradual decrescente para áreas até 1000 m de cursos de água, a fim de se evitar a proposição de empreendimentos em áreas que possam impactar os recursos hídricos.

***Apesar de o Conama ter revogado em 2010 a Resolução Conama 13/1990 (BRASIL, 1990) que determinava a zona de amortecimento das UCs sem plano de manejo em 10 mil metros, e determinado o novo valor de 3.000 metros, o estudo adotou um buffer de 10.000 m considerando a importância dessas zonas para o ecossistema das UCs.

Fonte: Souza et al. (2020).

A atribuição de pesos e notas às variáveis que serão analisadas, faz parte da metodologia de algumas técnicas de análise multicritério.

Segundo alguns autores, o método Knowledge Driven, frequentemente aplicado na definição de pesos e notas a partir do conhecimento de especialista, insere subjetividade nas análises. Li (2020) argumenta que o aprendizado de máquina, mais especificamente aprendizado profundo, permitem uma abordagem orientada aos dados, ou Data Driven, de forma que dados geoespaciais massivos, que são difíceis de manusear usando métodos tradicionais de análise espacial, podem ser analisados, minerados e visualizados (Li, 2020).

Os modelos de aprendizados de máquina podem ser separados em dois grandes grupos, de uma forma simplificada: (i) aprendizado supervisionado, onde os dados de treinamento contêm uma amostra de dados com a classificação correta atribuída, (ii) aprendizado não supervisionado, onde os modelos identificam padrões ocultos em um conjunto de dados de entrada não rotulados (Sathya, R., Nivas, J. & Abraham, A., 2013).

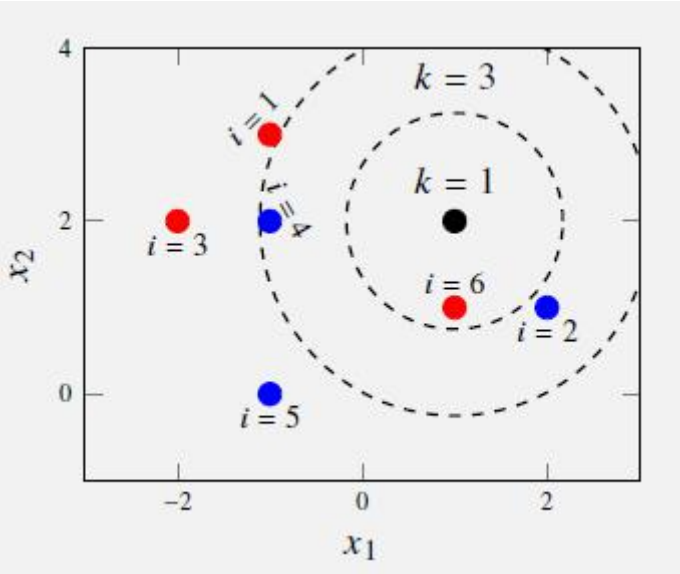
No aprendizado supervisionado utiliza-se um modelo ou método matemático para predizer o valor de uma variável y a partir de um conjunto de variáveis x. O processo de obtenção desse modelo matemático utiliza um conjunto de dados de treinamento onde as variáveis de entrada x estão associadas uma variável de saída y conhecida. A partir dos dados de treinamento é possível ajustar/generalizar um modelo que posteriormente será aplicado aos dados de teste (não conhecidos durante o processo de treinamento) (Lindholm et al., 2021).

Dentro do aprendizado supervisionado temos a classificação, onde a variável que se quer predizer é uma determinada classe ou rótulo, tendo apenas um número finito de possíveis valores. E a regressão, onde a variável que se quer predizer é um valor numérico.

(CC BY 4.0) | ISSN 2675-8539 | DOI: http://dx.doi.org/10.52076/eacad-v3i3.267

Um método de classificação supervisionada bastante utilizado é o K-vizinhos ou K-NN (K-Nearest Neighbors). Este método é baseado na ideia de que a distância entre os atributos da variável que se quer classificar é menor em relação ao conjunto de atributos da classe à qual ela pertence. De maneira geral o método segue os passos demonstrados na Figura 4, são consideradas apenas duas classes, vermelha e azul para facilitar o entendimento.

Figura 4 - Exemplo de classificação utilizando K-vizinhos.



Fonte: Lindholm et al. (2021)

1. Calcula-se a distância entre cada elemento do conjunto de dados de treinamento e dos atributos da variável que se quer classificar;
2. Determina-se o K-vizinhos mais próximos;
3. Determine-se o número de vizinhos de cada classe (valor de K);
4. Classifica-se a variável como pertencente à classe que possui o maior número de vizinhos.

Alguns pontos importantes que precisam ser considerados para utilização desse método são:

- É preciso definir a métrica de distância mais adequada de acordo o problema a ser resolvido, por exemplo, distância Euclidiana, Minkowski, Cosseno ou Person.
- No caso de utilização da distância Euclidiana é necessário normalizar a base de dados.
- Identificar o valor ótimo de K é fundamental, uma vez que valores de K muito pequenos ou muito grandes geram classificações pouco precisas.

Outros exemplos de métodos de classificação supervisionada são o Naive Bayes, Random Forest e as Redes Neurais. Para uma análise comparativa desses e de outros métodos consultar (Stephens D. & Diesing M., 2014).

Como exemplo de aprendizado não supervisionado podemos considerar o método K-means. Este é um método amplamente utilização para agrupamento de variáveis. O método K-means é numérico, não supervisionado e gera um número específico de clusters separados e não hierárquicos (Wang et al., 2020).

Em sua versão mais clássica o K-means começa inicializando um conjunto de K centroides, sendo que uma forma comum é escolha aleatória de k objetos do conjunto de dados para representar os centroides. Em seguida cada objeto do conjunto é associado ao cluster com o centroide mais próximo. Então os centroides são recalculados e o processo é repetido até que os

centroides não se alterem mais (Faceli et al, 2011).

Assim como no K-vizinhos, no K-means é preciso determinar a métrica de distância utilizada. O algoritmo é sensível à escolha inicial dos centroides (Faceli et al, 2011).

A Análise de Impacto Ambiental é parte fundamental dos processos de licenciamento ambiental e sua importância é amplamente reconhecida. A definição da localização geográfica de um empreendimento é um ponto chave para sua viabilização já que, além dos aspectos físicos e socioambientais, é considerara um recurso estratégico para as empresas. Porém determinar a melhor localização geográfica é um processo complexo, envolve grande número de variáveis, critérios dependentes de característica específicas do negócio, os quais muitas vezes são conflitantes.

Essa complexidade faz com que seja necessária a utilização de ferramentas de geoprocessamento, combinadas com técnicas de Análise de Multicritério para a avaliação das alternativas locacionais, porém estas técnicas são conhecidas como Knowledge Driven, ou seja, muito dependentes do conhecimento de especialistas, o que por sua vez, pressupõem alguma subjetividade.

Por outro lado, existem diversos algoritmo de aprendizado de máquina que podem representar uma abordagem mais orientada aos dados ou Data Driven. Através de modelos de classificação, regressão e agrupamento é possível analisar e visualizar grandes quantidades de dados geoespaciais. Isto por sua vez, pode tornar o processo mais automatizado e menos subjetivo.

3. Metodologia

Souza et al. (2020) em seu trabalho Otimização locacional de portos secos para fomentar o desenvolvimento regional sustentável, realizam uma avaliação da distribuição dos portos secos do Estado de Minas Gerais e identificam os melhores locais para uma nova instalação. A abordagem integrou duas metodologias distintas; a Análise Multicritério Espacial e o Problema de Localização-Alocação. Na Análise Multicritério Espacial foram atribuídos pesos às variáveis sociais, ambientais e econômicas. O Problema de Localização-Alocação adotou o resultado da análise multicritério como demanda ponderada e aplicou uma meta heurística para a solução do problema de otimização.

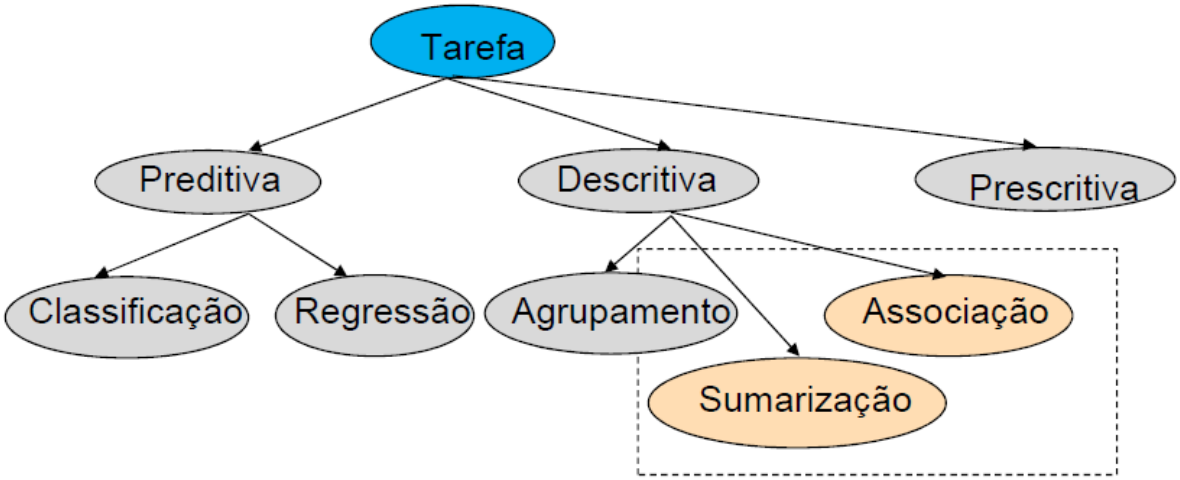
Para realização da análise multicritério é necessário tomar a opinião de especialistas, definir as variáveis mais relevantes no problema e atribuir pesos as essas variáveis. Da mesma forma no Problema de Localização-Alocação é necessário estabelecer uma série de regras, com base em conhecimentos de especialistas, para a construção da rede.

Estre trabalho tem como objetivo testar a aplicação de abordagens de aprendizado de máquina, de forma a tentar tornar a identificação das melhores localizações para a instalação de portos secos, mais automatizada, menos dependente da opinião de especialistas e mais orientada a dados. As variáveis ou features utilizadas no trabalho foram definidas com base nas descritas no artigo citado.

As tarefas de aprendizado de máquina podem ser separadas entre tarefas preditivas, tarefas descritivas e tarefas prescritivas, como apresentado na Figura 5, as duas tarefas destacadas são variações do agrupamento.

(CC BY 4.0) | ISSN 2675-8539 | DOI: http://dx.doi.org/10.52076/eacad-v3i3.267

Figura 5 - Tarefas de aprendizado de máquina.



Fonte: Apostila de Aprendizado de Máquina - CeMEAI-USP.

As tarefas preditivas buscam identificar um modelo capaz de predizer o rótulo a partir de valores de atributos preditivos ou dados de exemplo. Para criar tal modelo são utilizados algoritmos de aprendizado de máquina que, por sua vez são treinados (aprendem) através de um conjunto de dados de treinamento e são avaliados através de um conjunto de dados de teste.

As tarefas descritivas também buscam modelos apesar de no geral, não necessitarem de uma fase de treinamento. Alguns exemplos tarefas prescritivas são; agrupamento (clustering) cujo objetivo é organizar objetos não rotulados em grupos, de acordo com uma mediada de proximidade; sumarização cujo objetivo é encontrar uma descrição simples e resumida de um conjunto de dados; e associação de itens frequentes, onde dado um conjunto de itens e uma base de dados de transações, o objetivo é encontrar um conjunto de regras, nas várias transações, que associem a presença de um item à presença de outros itens.

Existem ainda as tarefas prescritivas que ao contrário das tarefas preditivas, predizem a entrada que é necessária para produzir uma saída desejada. Um exemplo de tarefa prescritiva é o controle de robôs, que prediz a entrada de controle para que uma trajetória seja seguida.

Este trabalho vai se concentrar na regressão, que faz parte do grupo de tarefas preditivas e é utilizada em diversas aplicações. Seu objetivo é aprender uma função capaz de associar dados de atributos a um valor real.

No processo de regressão utiliza-se um conjunto de dados para predizer o valor de uma variável de saída. O modelo de regressão é dado por:

$$\boldsymbol{y = f(X, \theta) + \epsilon}$$

Onde:

ϵ é o erro, representando o que não poder ser capturado pelo modelo.

Um dos modelos mais utilizados é o modelo de regressão linear que é capaz de predizer dados desconhecidos a partir de um modelo treinado, além de poder determinar a importância de cada variável na previsão.

O modelo de regressão linear multivariada e dado por:

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_d X_d}$$

Onde:

Y = variável resposta (que se quer predizer)

β_0,β_1,β_2,β_d = parâmetros

X_1,X_2,X_d = variáveis independentes

Para estimar os coeficientes utiliza-se o conjunto de dados de treinamento, obtendo-se o modelo ajustado, representado por:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_{1\,x1} + \hat{\beta}_{2\,x2} + \cdots \hat{\beta}_{d\,xd}$$

O erro é a diferença entre o valor real e o valor predito no conjunto de dados de treinamento. Na regressão múltipla busca-se encontrar os valores dos parâmetros β_i de forma a minimizar o erro. Minimizar função de custo j (w_0,w_1,w_2… w_d)

$$jw = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Algoritmos de regressão podem induzir modelos sobre-ajustados causando o chamado overfitting. O overfitting ocorre quando o modelo decora os valores dos parâmetros e dessa forma não apresenta boa performance na predição de novos conjuntos de dados. Alguns exemplos de fatores que podem causar o overfitting são a baixa proporção do número de exemplos em relação ao número de atributos preditivos e uma distribuição dos dados é muito complexa. A incorporação de um termo de regularização para restringir o valor dos parâmetros, restringe a complexidade do modelo induzido, podendo evitar o overfitting.

A técnicas de regularização adicionam um termo de penalidade à função de custo quando coeficientes têm valores elevados. Duas técnicas de regularização muito utilizadas são a LASSO (least absolute shrinkage and selection operator), cuja penalidade é aplicada à soma dos valores absolutos dos coeficientes, e a Ridge, cuja penalidade é aplicada à soma dos quadrados dos coeficientes.

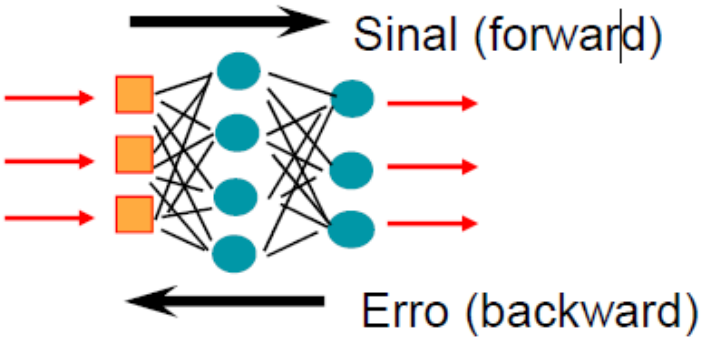
Neste trabalho serão avaliados os resultados da regressão considerando as regularizações LASSO e Ridge.

Outra forma de se induzir modelos de regressão é através das Redes Neurais. As Redes Neurais Artificias (RNA) são sistemas distribuídos inspirados no cérebro humano. As RNAs possuem uma arquitetura baseada em unidades de processamento (neurônios), conexões ou sinapses e topologia que é a forma que os neurônios e suas conexões assumem.

A Mult-Layer Perceptron (MLP) e umas das arquiteturas de RNA mais utilizadas. Possui uma ou mais camadas intermediárias de neurônios. As MLPs funcionam através de pares entrada-saída, onde cada vetor de entrada é associado a uma saída desejada. Estas redes procuram associar os valores de um vetor de entrada a valores de saída, através do ajuste dos pesos e minimização dos erros. O treinamento é feito em duas fases, cada uma percorrendo a rede em um sentido, fase forward na direção da saída e fase backward na direção contrária conforme Figura 6, note que com apenas 3 camadas o número de conexões já é bastante alto.

(CC BY 4.0) | ISSN 2675-8539 | DOI: http://dx.doi.org/10.52076/eacad-v3i3.267

Figura 6 - Sentido das fases de treinamento em uma rede MLP.



Fonte: Apostila de Aprendizado de Máquina - CeMEAI-USP.

Para medir o desempenho de um modelo pode-se considerar sua capacidade preditiva, seu custo computacional, ou sua interpretabilidade. Nos processos de regressão a capacidade preditiva dos modelos é medida pela diferença entre o valor predito e o valor real. Alguns exemplos de medidas de desempenho são:

Soma dos quadrados dos erros (SSE)

$$\boldsymbol{SSE} = \sum_{i=1}^{n} (\boldsymbol{y}^i - \boldsymbol{f}(\boldsymbol{x}^i))^2$$

Erro quadrático médio (MSE)

$$\boldsymbol{MSE} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{i=1}^{n} (\boldsymbol{y}^i - \boldsymbol{f}(\boldsymbol{x}^i))^2$$

Erro absoluto médio (MAE)

$$\boldsymbol{MAE} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{i=1}^{n} |\boldsymbol{y}^i - \boldsymbol{f}(\boldsymbol{x}^i)|$$

Coeficiente de determinação (R2)

$$\boldsymbol{R^2} = \boldsymbol{1} - \frac{\boldsymbol{MSE}}{\boldsymbol{Var(y)}}$$

O Coeficiente de determinação (R2) é uma versão padronizada do MSE, sendo mais fácil de interpretar por ter mesma unidade de valor que y.

Neste trabalho serão definidas duas métricas distintas para avaliar a performance dos modelos gerados.

Portos secos podem ser definidos como terminais intermodais terrestres conectados aos portos marítimos através de serviços de transporte de alta capacidade (Nguyen; Notteboom, 2016; Roso; Woxenius; Lumsden, 2009 apud Souza at al, 2020).

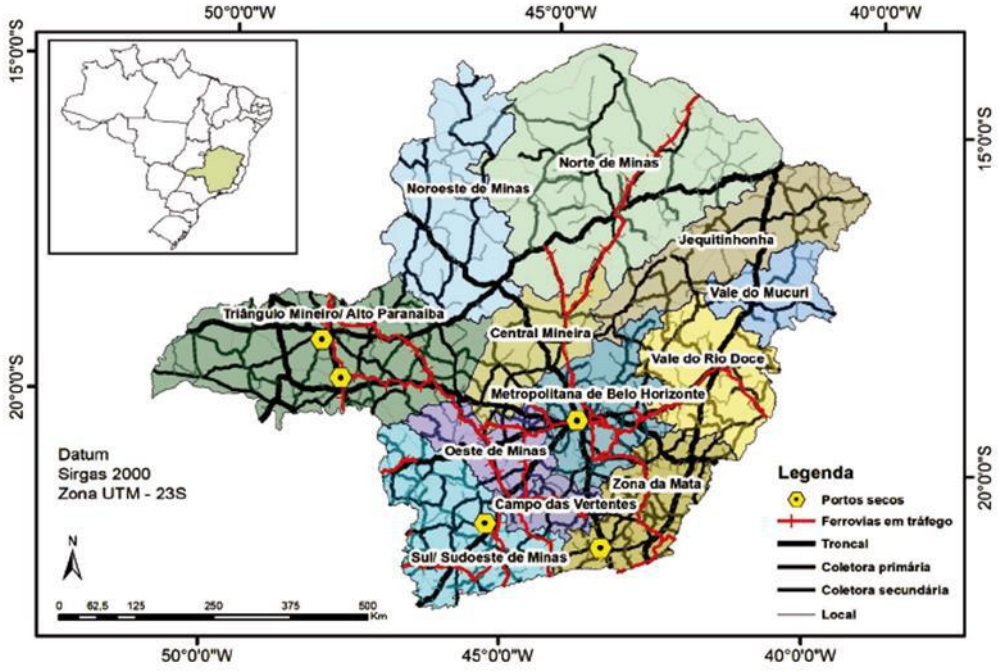
Minas Gerais possui grande importância logística e econômica no cenário brasileiro, com exportação de produtos como minério de ferro, café, soja, dentre outros. O fato de o Estado não ter acesso a portos marítimos representa condição ideal para instalação de portos secos.

O estado de MG apresenta ainda grande extensão territorial, elevado número de municípios, alta heterogeneidade biofísica, socioeconômica e cultural. Estas características tornam a escolha da melhor localização para a instalação de um porto seco extremamente complexa.

Por fim, o estado de MG possui a malha rodoviária mais extensa e a segunda malha ferroviária mais extensa do Brasil.

A Figura 7 apresenta a área de estudo, com suas mesorregiões, ferrovias em tráfego, principais rodovias (federais e estaduais) e a localização dos portos secos existentes, é possível verificar uma concentração de portos secos nas regiões ao sul do estado.

Figura 7 - Mesorregiões de Minas Gerais com os portos secos, ferrovias em tráfego e as principais rodovias existentes.



Fonte: Souza at al. (2020).

A criação da base de dados para o estudo se iniciou com uma busca pelas fontes citadas no artigo utilizado como referência. Alguns dos dados são disponibilizados apenas mediante solicitação, outros foram obtidos em fontes diferentes das citadas no artigo. O Quadro 1 apresente as fontes dos dados, procurou-se manter a mesma estrutura adotada no estudo de referência.

Quadro 1 - Fontes de dados.

Grupo	Variável	Fonte	Data	Formato Dado	Método	Fonte possível
Econômico	Valor FOB exp	MDIC	2015	Tabular	Interpolação (IDW)	https://www.gov.br/produtividade-e-comercio-exterior/pt-br/acesso-a-informacao/dados-abertos
	Valor FOB imp	MDIC	2015	Tabular	Interpolação (IDW)	https://www.gov.br/produtividade-e-comercio-exterior/pt-br/acesso-a-informacao/dados-abertos
	nº postos de trabalho	RAIS	2010	Tabular	Interpolação (IDW)	https://portalfat.mte.gov.br/relacao-anual-de-informacoes-sociais-rais/
	PIB	FGV	2015	Tabular	Interpolação (IDW)	https://dados.gov.br/dataset/cgeo_vw_pib_percapita
Competição	Distância PS existente	PNLT 2010	2010	Vetor Ponto	Distância Euclidiana	https://www.gov.br/infraestrutura/pt-br/assuntos/transporte-terrestre/base-de-dados-georreferenciada
Acessibilidade	Dist. Rod. Troncal	DEER/Setop	2015	Vetor Linha	Dist. Euclidiana	https://drive.google.com/u/0/uc?id=1YbmhT21Q_myh65dW5hZaVpdwtmC5HszW&export=download
	Dist. Rod. Principal	DEER/Setop	2015	Vetor Linha	Dist. Euclidiana	https://drive.google.com/u/0/uc?id=1YbmhT21Q_myh65dW5hZaVpdwtmC5HszW&export=download
	Distância ferrovias	PNLT 2010	2010		Dist. Euclidiana	https://www.gov.br/infraestrutura/pt-br/assuntos/transporte-terrestre/base-de-dados-georreferenciada
Social	nº de não ocupados	IBGE	2010	Tabular	Interpolação (IDW)	https://sidra.ibge.gov.br/Tabela/3580
	Índice Social	Atlas Brasil	2010	Tabular	Interpolação (IDW)	http://www.atlasbrasil.org.br/consulta/planilha
Ambiental	Dist. área urbana	Embrapa Territorial	2015	Vector - Polígono	Dist. Euclidiana	https://www.ibge.gov.br/geociencias/cartas-e-mapas/redes-geograficas/15789-areas-urbanizadas.html?=&t=downloads
	Dist. Hidrografia	IBGE	2015	Vetor - Linha	Dist. Euclidiana	http://www.igam.mg.gov.br/banco-de-noticias/1-ultimas-noticias/1312-hidrografia
	Dist. Ucs	ICMBio/Semad	2015	Vector - Polígono	Dist. Euclidiana	http://mapas.mma.gov.br/i3geo/datadownload.htm

Fonte: Autores.

Apesar de todas as fontes de dados serem púbicas, durante a execução do trabalho não foi possível acessar os dados de nº postos de trabalho e nº de não ocupados.

A partir da obtenção dos dados o próximo passo foi o tratamento espacial, contemplando o recorte dos dados vetoriais para o estado de MG e correção de erros de geometria. Os valores tabulares foram associados às coordenadas das sedes municipais. Para dar suporte à estruturação dos dados foi utilizada a ferramenta QGIS.

A partir da estruturação inicial dos dados foi necessária a geração de superfícies. Essas superfícies possuem um formato matricial, o que possibilita a conversão doados para o formato de data frames, que são tratados pelo Python.

A área representada por cada pixel, ou tamanho do pixel. foi definida em 500 metros, de forma a viabilizar o processamento local das informações a manter o Padrão de Exatidão Cartográfica (mais informações em http://www.planalto.gov.br/ccivil_03/decreto/1980-1989/d89817.htm).

Foram utilizadas técnicas de interpolação para gerar superfícies a partir de dados tabulares, como os valores Valor FOB exp, Valor FOB imp e PIB. A técnica utilizada foi a Ponderação pelo Inverso da Distância (IDW). No IDW a ponderação é atribuída a pontos amostrais através da utilização de um coeficiente de ponderação que controla como a influência da ponderação irá diminuir à medida que a distância a partir do ponto desconhecido aumenta. Como os valores para parametrização não foram citados no artigo, foram mantidos os valores padrão, adotados pela ferramenta QGIS.

Para os dados vetoriais Dist. Rod. Troncal, Dist. Rod. Principal, Distância ferrovias, Dist. área urbana, Dist. Hidrografia,

Dist. Ucs, foi utilizada a Distância Euclidiana, onde cada pixel dos dados de saída representa a distância do pixel mais próximo dos dados de entrada.

Conforme definido no artigo, área de influência para as áreas urbanas foi limitada a 5 km. Para a hidrografia foi definida uma oneração gradual decrescente para áreas até 1000 m de cursos de água, para evitar a localização de empreendimentos em áreas que possam impactar os recursos hídricos. Para as Unidades de Conservação (UCs) o estudo adotou um buffer de 10.000 m considerando a importância dessas zonas para o ecossistema das UCs.

O estudo utilizado como referência propõem ainda a criação da variável Índice Social (IS) definida conforme descrito abaixo:

"O IS foi proposto para ponderar, além da demanda por serviços portuários, também o potencial de impacto socioeconômico regional do PS. O IS é proporcional à população e inversamente proporcional ao quadrado do IDH conforme mostra a Equação 1" (Souza at al., 2020)

O cálculo do IS é dado pela fórmula IS = Pop/IDH2, onde Pop é a população do munícipio e IDH é o índice de desenvolvimento social.

As variáveis avaliadas na determinação da melhor localização para a instalação do porto seco possuem diferentes ordens de grandeza sendo necessário normalizá-las. As variáveis também contribuem de forma diferente para terminação do custo de oportunidade. Por exemplo, quando se considera uma área de preservação, quanto mais distante do porto seco mais atrativa é a localização, já quando se considera uma ferrovia, quanto mais próxima do porto seco, maior a atratividade. Sendo assim foi necessário estabelecer uma faixa de valores de 0 a 10, onde zero (0) representa ausência de influência da variável, um (1) representa o maior custo de oportunidade e 10 o menor custo de oportunidade. Quanto menor o custo de oportunidade mais atrativa é a localização e consequentemente, quanto maior o custo, menos atrativa é a localização.

O Quadro 2 apresenta a interpretação das variáveis e a forma como seu valor (resultado da interpolação) contribui para o custo de oportunidade, em algumas variáveis a relação inversamente proporcional entre o resultado a interpolação e o custo de oportunidade gera uma atratividade alta, em outras essa lógica se inverte

Quadro 2 - Interpretação das variáveis.

Variáveis	Interpretação	Valor	Custo de Oportunidade	Atratividade
Valor FOB exp	Valor de exportações em dolar	Alto	Baixo	Alta
Valor FOB imp	Valor de importações em dolar	Alto	Baixo	Alta
PIB	Produto interno bruto do município	Alto	Baixo	Alta
Distância PS existente	Distância da localização em relação a portos secos existentes	Alto	Baixo	Alta
Dist. Rod. Troncal	Distância da localização em relação a rodovias troncais	Alto	Alto	Baixa
Dist. Rod. Principal	Distância da localização em relação a rodovias principais	Alto	Alto	Baixa
Distância ferrovias	Distância da localização em relação a ferrovias	Alto	Alto	Baixa
Índice Social	Ponderação do impacto social da instalação do PS	Alto	Baixo	Alta
Dist. área urbana	Distância da localização em relação a área urbana	Alto	Alto	Baixa
Dist. Hidrografia	Distância da localização em relação a área de preservação	Alto	Baixo	Alta
Dist. Ucs	Distância da localização em relação a unidades de conservação	Alto	Baixo	Alta

Fonte: Autores.

Portos secos podem ser vistos como extensões de zonas de influência portuária, facilitando a movimentação de cargas e contribuindo para melhorar a agilidade.

Minas Gerais, por suas características físicas e sociais tem grande potencial para instalação de portos secos. Porém, definir a melhor localização para uma nova instalação representa um grande desafio, principalmente quando consideramos a extensão, infraestrutura e heterogeneidade do Estado.

Em razão dessa complexidade foi utilizado como referência o estudo realizado por Souza at al, (2020) e as variáveis definidas como de maior importância na Análise Multicritério. Essas variáveis consideram aspectos econômicos, de competição, de acessibilidade, sociais e ambientais.

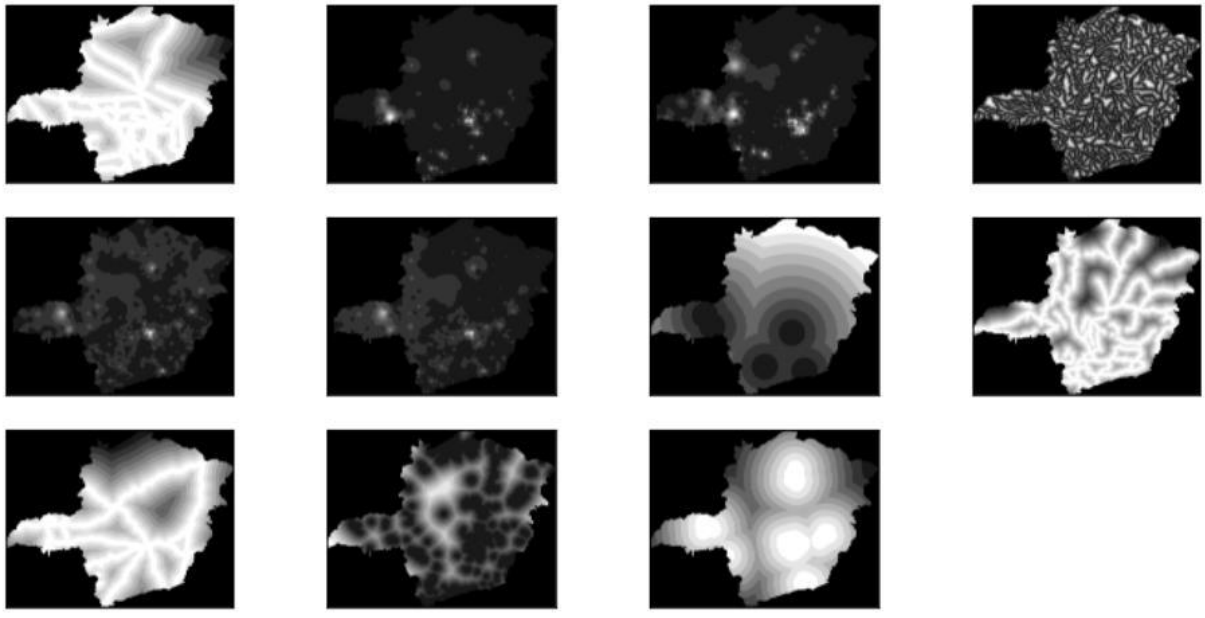
Os dados que compõem cada aspecto avaliado foram obtidos em fontes oficiais, em sua maioria, a partir das referências citadas no artigo.

Após a obtenção dos dados foi necessário realizar seu processamento, correções e geração de uma base integrada, com o suporte da ferramenta QGIS.

Como os dados possuem ordens de grandezas e interpretações diferentes, o próximo passo foi propor uma normalização, de forma que os valores obtidos pelos processos de interpolação (IDW) e Distâncias Euclidianas, pudessem ser avaliados corretamente pelos modelos de regressão.

Por fim, foram geradas superfícies no formato matricial ou raster, que posteriormente foram convertidas para a biblioteca geopanda do Python. A Figura 8 apresenta as variáveis em formato raster, as áreas mais claras estão associadas ao menor custo de atratividade ou maior atratividade.

Figura 8 - Variáveis em formato raster.



Fonte: Autores.

4. Resultados e Discussão

O estudo utilizado como referência apresenta uma lista dos municípios com o menor custo de oportunidade, ou seja, a maior atratividade para a implantação de um porto seco. O Quadro 3 a seguir, apresenta o resultado obtido pelo estudo na forma a lista ordenada, crescente.

Quadro 3 - Lista de município ordenada por custo de oportunidade.

Análise multicritério (Artigo)	**Custo**
BETIM	2,86
MONTES CLAROS	3,07
ARAXA	3,22
BELO HORIZONTE	3,25
SETE LAGOAS	3,30
UBERABA	3,37
ITABIRA	3,75
PARACATU	3,75
UBERLÂNDIA	3,78
SANTANA DO PARAÍSO	3,83
ITABIRITO	3,96
ARAGUARI	3,97

Fonte: Autores.

Conforme destacado anteriormente, foi necessário estabelecer uma faixa de valores de 0 a 10 para normalizar as variáveis e representar a atratividade da localidade para instalação do porto seco. Como técnicas de regressão pressupõem a existência de valores rotulados, optou-se por definir uma nota para os municípios listados no estudo. Para isso foi aplica a equação geral da reta (y = a x + b) aos custos e obtida a nota do município. O Quadro 4 apresenta a lista de municípios com suas respectivas notas, mantendo-se a ordem original do estudo.

Quadro 4 - Lista de municípios e notas.

Análise multicritério (Artigo)	**Custo**	**Nota**
BETIM	2,86	10,03
MONTES CLAROS	3,07	8,14
ARAXA	3,22	6,79
BELO HORIZONTE	3,25	6,52
SETE LAGOAS	3,30	6,07
UBERABA	3,37	5,44
ITABIRA	3,75	2,02
PARACATU	3,75	2,02
UBERLÂNDIA	3,78	1,75
SANTANA DO PARAÍSO	3,83	1,30
ITABIRITO	3,96	0,13
ARAGUARI	3,97	0,04

Fonte: Autores.

É importante salientar que apesar da lista conter apenas 12 munícipios o volume de dados rotulados é bem maior, uma vez que os dados estão em formato matricial ou raster, e dessa forma, foram atribuídos valores a todos os pixels em um buffer de 500 m2 ao redor dos municípios.

Inicialmente os dados foram separados em exemplos e rótulos e posteriormente entre dados de treinamento e teste, na proporção de 70% para treinamento e 30% para teste.

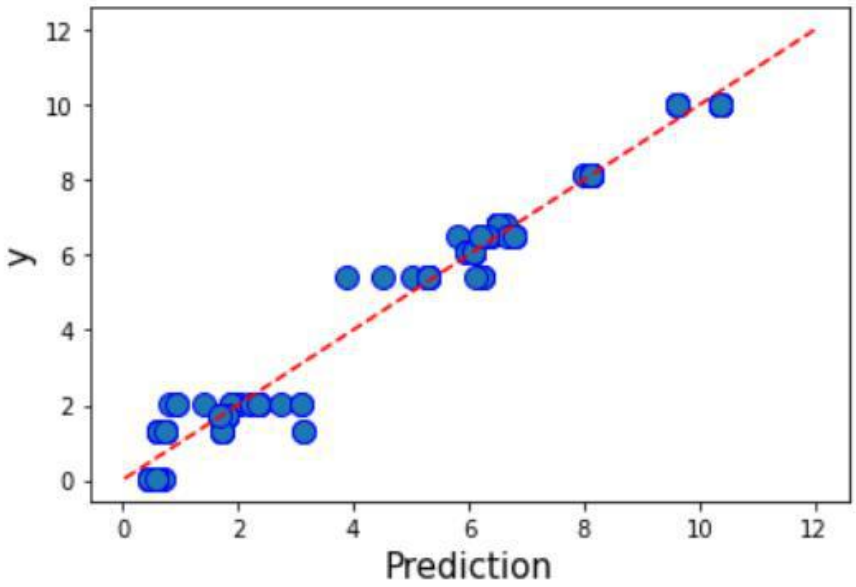
(CC BY 4.0) | ISSN 2675-8539 | DOI: http://dx.doi.org/10.52076/eacad-v3i3.267

Foram aplicados três algoritmos de regressão linear multivariada, disponíveis na biblioteca sklearn do python.

Foram consideradas onze (11) variáveis para a geração do modelo. Como forma de visualizar a precisão da predição optou-se pela geração de gráficos com os valores reais versus os valores preditos. Os resultados são apresentados a seguir.

A Figura 9 apresenta o resultado do modelo de regressão linear múltipla – LinearRegression. A análise do gráfico permite concluir que o modelo alcançou um bom desempenho na predição, com a maioria dos prontos bem próximos da reta, não sendo possível identificar a presença de muitos outliers.

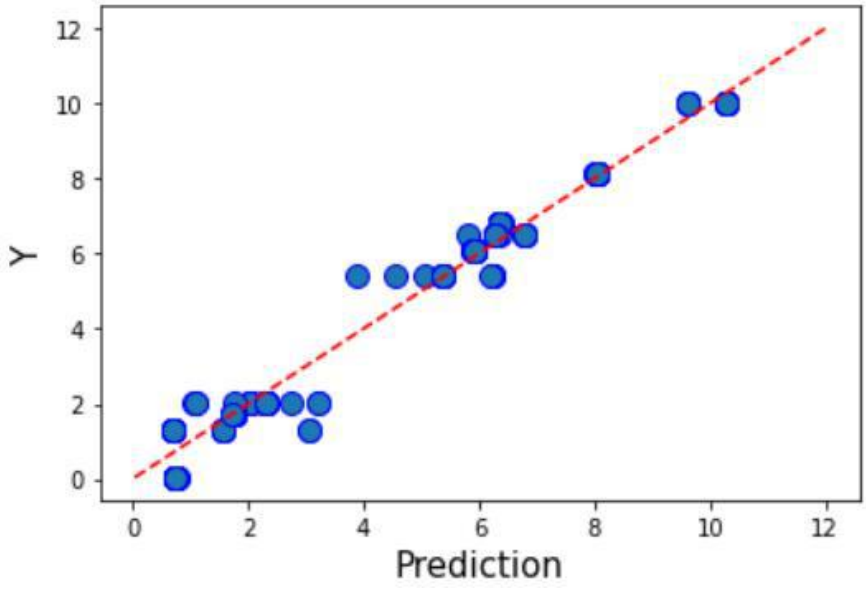
Figura 9 - Resultado modelo de regressão linear.



Fonte: Autores.

A Figura 10 apresenta o resultado do modelo de regressão Ridge – Ridge(). Assim como no modelo de regressão linear, a análise do gráfico permite concluir que o modelo alcançou um bom desempenho na predição, com a maioria dos pontos bem próximos da reta, não sendo possível identifica a presença de muitos outliers. O modelo foi testado com diferentes valores de alpha, não apresentando alterações substanciais nos resultados da predição.

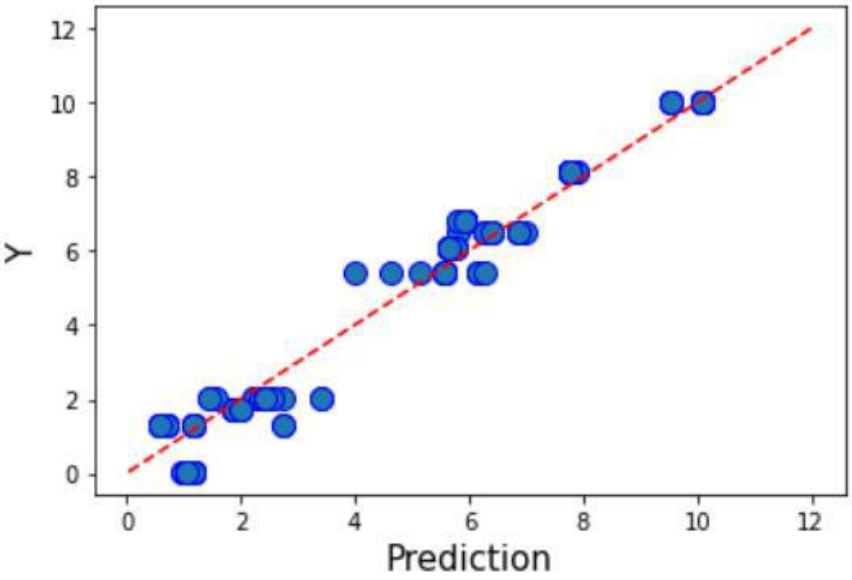
Figura 10 - Resultado da regressão Ridge.



Fonte: Autores.

A Figura 11 apresenta o resultado do modelo de regressão LASSO – Lasso(). A análise do gráfico permite concluir que modelo alcançou um bom desempenho na predição, com a maioria dos pontos próximos da reta, não sendo possível identifica a presença de muitos outliers. Assim como na regressão Ridge o modelo LASSO foi testado com diferentes valores de alpha, não apresentando diferenças substanciais nos resultados da predição.

Figura 11 - Resultado regressão LASSO.



Fonte: Autores.

Além dos gráficos dos valores reais e valores preditos, foram coletadas métricas de desempenho para cada um dos modelos. As métricas coletadas foram Erro Quadrático Médio (MSE) e o Coeficiente de Determinação (R2). O Quadro 5 apresenta as métricas de desempenho para cada um dos modelos de regressão. Assim que como na análise gráfica os valores das métrica coletadas demonstram que os modelos apresentaram boa performance. Observa-se que os resultados de MSE e R2 foram

os mesmos para os modelos de regressão linear e regressão Ridge. Já o modelo de regressão LASSO apresentou um resultado um pouco que pior que os dois anteriores.

Quadro 5 - Métricas de desempenho dos modelos de regressão.

MODELO	MSE	R2
Regressão Linear	0,2438	0,9762
Regressão Ridge	0,2438	0,9762
Regressão LASSO	0,4068	0,9603

Fonte: Autores.

Além dos modelos gerados pela LinearRegression(), Ridge() e LASSO(), foi testado um modelo gerado por uma rede neural sequencial. Como não se dispunha de uma grande quantidade de dados de treinamento, optou-se por uma arquitetura mais rasa. A arquitetura da rede é apresentada na Figura 12, contendo apenas três camadas densas.

Figura 12 - Arquitetura da rede neural sequencial

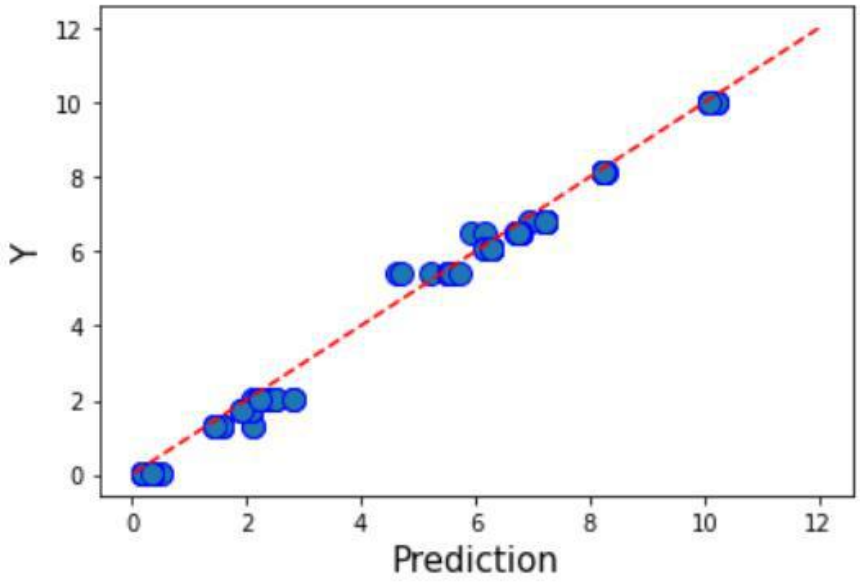
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                768

 dense_1 (Dense)             (None, 64)                4160

 dense_2 (Dense)             (None, 1)                 65

=================================================================
Total params: 4,993
Trainable params: 4,993
Non-trainable params: 0
_________________________________________________________________
```

Fonte: Autores.

A arquitetura apresentada foi testada com diferentes tamanhos de lotes e com diferentes quantidades de épocas não apresentando diferenças substanciais nos resultados da predição.

A Figura 13 apresenta o resultado do modelo de regressão com a rede neural sequencial. Assim como nos modelos de regressão, a análise gráfica permite concluir que modelo alcançou um bom desempenho na predição com a maioria dos prontos bem próximos da reta, não sendo possível notar a presença de muitos outliers.

Figura 13 - Resultados da regressão com rede neural sequencial.



Fonte: Autores.

Porém as métricas de MSE e R2 mostram um desempenho superior da rede neural. O Quadro 6 apresenta as métricas de desempenho para cada um dos modelos de regressão, onde verificam-se valores superiores, tanto de MSE quanto de R2, para a rede neural.

Quadro 6 - Métrica de desempenho dos modelos de regressão e da rede neural.

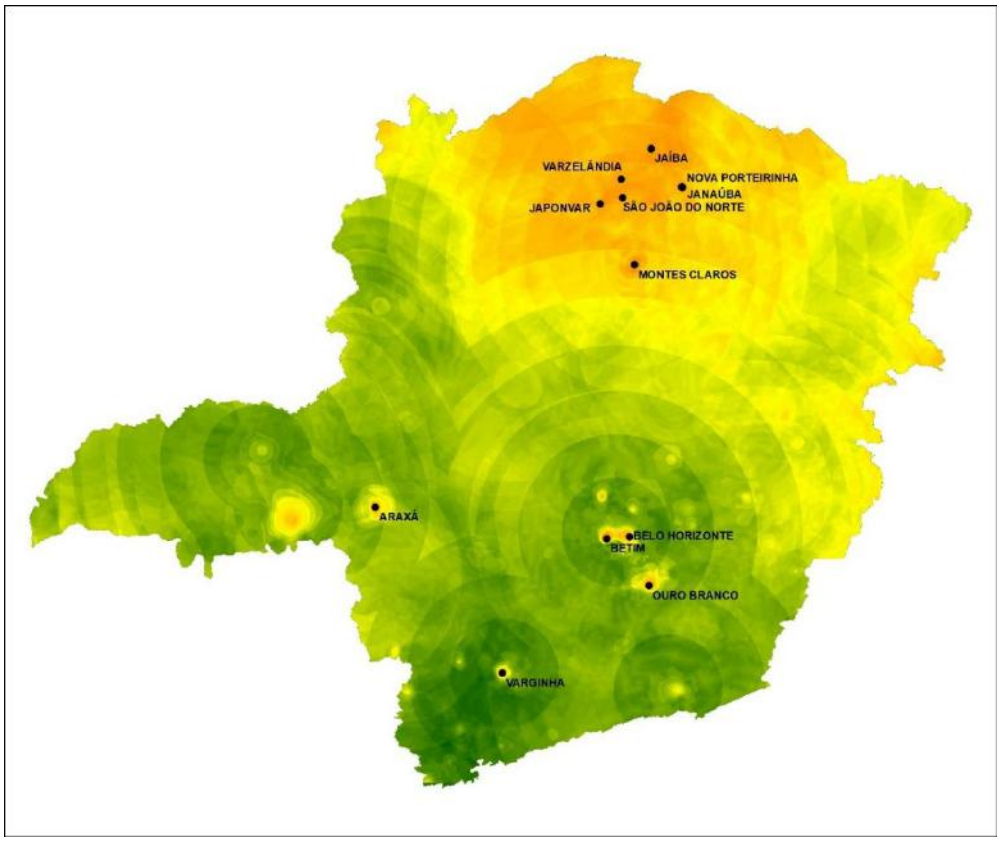
MODELO	MSE	R2
Regressão Linear	0,2438	0,9762
Regressão Ridge	0,2438	0,9762
Regressão LASSO	0,4068	0,9603
Rede Neural Sequencial	0,0885	0,9914

Fonte: Autores.

Como o modelo gerado pela rede neural apresentou a melhor performance, ele foi o escolhido para ser aplicado à base de dados, para definição das áreas de maior atratividade. Após a aplicação do modelo foi criado um arquivo do tipo csv, para visualização dos resultados no software GIS.

A Figura 14 apresenta as áreas de maior atratividade obtidas pela rede neural, na forma de um mapa de calor. É possível verificar que na região central e na região sul do estado existe uma grande similaridade entre os municípios identificados através do modelo de aprendizado de máquina com aqueles identificados no artigo de referência. Na região norte o modelo identificou outras áreas de alta atratividade para a instalação de portos secos.

Figura 14 - Mapa de calor das áreas de maior atratividade.



Fonte: Autores.

A aplicação de modelos de regressão para predizer as áreas de maior atratividade apresentou resultados bastante satisfatórios. Foi possível verificar que uma boa acurácia em relação aos dados de testes, além de boa generalização para o conjunto de dados novos.

As diferenças entre as áreas identificadas pelo modelo e as identificadas pelo artigo podem se dever a diversos fatores, como por exemplo, à impossibilidade de acesso à base de dados utilizada no primeiro estudo, ao processo de normalização dos dados, à exclusão de algumas variáveis, à própria diferença de alcance das metodologias utilizadas.

De qualquer forma os resultados obtidos deixam claro que outras técnicas de aprendizado de máquina podem ser empregas, indicando resultados com grande potencial.

5. Conclusão

As técnicas de aprendizado de máquina têm ganhado cada vez mais relevância tanto nas universidades quanto nas empresas, e possuem um amplo espectro de aplicações. Particularmente as redes neurais têm obtido resultados cada vez mais impressionantes em tarefas preditivas complexas como Processamento Natural de Linguagem (NPL), classificação de imagens e detecção de fraudes.

Por sua vez os chamados dados geográficos representam um desafio especial para as tarefas de aprendizado de máquina, em função da forma particular com que precisam ser tratados, por normalmente apresentarem volumes muito grandes de dados, possuírem diferentes padrões de representação e diferente resoluções.

No caso da alternativa locacional de empreendimentos, aspectos diversos, e muitas vezes conflitantes, precisam ser considerados. Essas características tornam as análises extremamente complexas, dependentes de aspectos subjetivos e muito

difíceis de serem reproduzidas ou generalizadas.

Este trabalho teve como objetivo testar a aplicação de técnicas e algoritmos de aprendizado de máquina para a análise de alternativas locacionais. Como em diversos outros estudos a coleta e o processamento dos dados representou o maior esforço do projeto, necessitando de uma série de fases de estruturação, geolocalização, processamento e normalização.

Algumas das vantagens que merecem ser destacadas são; (1) a redução da subjetividade da análise, (2) mapeamento de clusters de áreas de maior atratividade ao invés de apenas municípios, (3) a possibilidade da aplicação dos modelos para outras áreas, reduzindo custos e aumentando a produtividade; (4) a proposição de uma abordagem inovadora para tratar a questão de geolocalização.

Entende-se que a abordagem desenvolvida representa um avanço para soluções de análise de alternativa locacional e espera-se aplicá-la, no curto prazo, para demandas reais das empresas.

Como sugestões para trabalhos futuros podemos considerar; (1) utilização de uma abordagem baseada em classificação, com a aplicação de algoritmos que implementam árvores de decisão; (2) aplicação do modelo de melhor desempenho para outras áreas de estudo; (3) teste de diferentes arquiteturas da rede neural como por exemplos as reder neurais convolucionais.

Referências

Carnasciali, A, M dos S. & Delazari, L. S. (2011). A Localização Geográfica como Recurso Organizacional: Utilização de Sistemas Especialistas para Subsidiar a Tomada de Decisão Locacional do Setor Bancário. RAC, Curitiba, 15(1), 103-125. https://doi.org/10.1590/S1415-65552011000100007

Carvalho, C., P., L., F., A. Aprendizado de Máquina. Apostila do Curso de Ciência de Dados. Universidade de São Paulo. ICMC São Carlos. CeMEAI.

Davis, C. & Fonseca F. (2001). Introdução aos Sistemas de Informação Geográficos, 2001. Apostila do Curso de Especialização em Geoprocessamento. Disponível em http://www.csr.ufmg.br/geoprocessamento/publicacoes/introducao%20aos%20SIG.pdf em 25/06/2021.

Decreto-Lei Nº 89.817, de 20 de junho de 1984. Estabelece as Instruções Reguladoras das Normas Técnicas da Cartografia Nacional. Recuperado de http://www.planalto.gov.br/ccivil_03/decreto/1980-1989/d89817.htm

Faceli K., Lorena, C. A., Gama, J. & Carvalho, F. L. A. (2011). Inteligência Artificial: Uma abordagem por aprendizado de máquina. Rio de Janeiro: LTC.

Fernandes, V. H. A., Cassiano, A. de M., Guimarães, S. C. T. & Almeida, R. R. M. (2017). Alternativas locacionais em Avaliação de Impacto Ambiental de rodovias mineiras. Desenvolv. Meio Ambiente, 43, Edição Especial: Avaliação de Impacto Ambiental, 73-90. Doi: 10.5380/dma. v43i0.54056

Furlanetto, T. (2012). Estudo de alternativas locacionais para viabilidade ambiental de empreendimentos: o caso de aeroporto de Ribeirão Preto – SP (Dissertação de mestrado). Escola de Engenharia, Ciências da Engenharia Ambiental, Universidade de São Paulo.

Li, W. (2020). GeoAI: Where machine learning and big data converge in GIScience. Journal of Spatial Information Science, 20, 71-77. Doi:10.5311/JOSIS.2020.20.658

Lindholm, A., Wahlstrom, N., Lindsten, F. & Schon, B. T. (2021). Machine Learning: A First Course for Engineers and Scientists. Draft version: April 30, 2021.

Mattos, N. A. S. (2018). Alternativas locacionais em estudos de impacto ambiental no Estado de São Paulo (Dissertação de mestrado). Escola de Artes, Ciência e Humanidades, Universidade de São Paulo.

Nikparvar, B.; Thill, J.-C. Machine Learning of Spatial Data. (2021). ISPRS Int. J. Geo-Inf., 10, 600. https://doi.org/10.3390/ijgi10090600. Disponível em https://www.mdpi.com/2220-9964/10/9/600/htm 14/02/2022

Rodrigues, A., F. (2019). Introdução a Ciências de Dados. Apostila do Curso de Ciência de Dados. Universidade de São Paulo. ICMC São Carlos. CeMEAI.

Sathya, R., Niva, J. & Abraham, A. (2013). Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. (IJARAI) International Journal of Advanced Research in Artificial Intelligence, 2(2), 34-38.

Souza de F. M., Pinto, G. H. P., Teixeira, A. B. R., Nascimento, L. O. de C. & Nóbrega A. de A. R. (2020). Otimização locacional de portos secos para fomentar o desenvolvimento regional sustentável. Sustainability in Debate, Brasília, 11(2), 223-237.

Spigolon, G. M. L. (2015). A otimização da rede de transporte de RSU baseada no uso do SIG e análise de decisão multicritério para localização de aterros sanitários. (Tese de Doutorado). Escola de Engenharia de São Carlos, Universidade de São Paulo.

Stephens, D. & Diesing, M. A. (2014). A Comparison of Supervised Classification Methods for the Prediction of Substrate Type Using Multibeam Acoustic and Legacy Grain-Size Data. PLOS ONE, 9(4): e93950. https://doi.org/10.1371/journal.pone.0093950

Xiaojia, W., Changyan, S., Sheng, X., Shanshan, Z., Weiqun, X. & Yuxiang, G. (2020) Study on the Location of Private Clinics Based on K-Means Clustering Method and an Integrated Evaluation Model. IEEE Access. Doi: 10.1109/ACCESS.2020.2967797.

Zambon, L. K. (2004). Localização de usinas termoelétricas utilizando sistemas de informação geográficas e métodos de decisão multicritério. (Tese de Douturado). Escola de Engenharia de São Carlos, Universidade de São Paulo.