



26º Encontro Nacional de Conservação Rodoviária (ENACOR) 49ª Reunião Anual de Pavimentação (RAPV)

METODOLOGIA INTEGRADA PARA IMPUTAÇÃO DE DADOS GEOTÉCNICOS RODOVIÁRIOS: COMBINAÇÃO DE MODELOS DE PREDIÇÃO

*Diego Vieira de Souza¹; Mariana Gonçalves da Silva²; Brunna Mikaelli Pinho da Silva³;
Francisco Junior Peixoto Dantas¹; Antonio Junior Alves Ribeiro³ & Suelly Helena de Araújo
Barroso²*

RESUMO

A previsão de dados geotécnicos tem sido essencial para a execução de projetos rodoviários e manutenção de rodovias, reduzindo custos e melhorando a sua qualidade. Inserindo técnicas estatísticas é possível prever as condições de pavimentação e ensaios do solo de maneira eficaz. Desse modo, o objetivo deste estudo é realizar a imputação de dados correlacionados às variáveis fundamentais no banco de dados oriundo de 47 projetos rodoviários do Departamento Nacional de Infraestrutura de Transportes (DNIT) e 10 projetos da Secretaria de Obras Públicas do Ceará (SOP/CE), onde o mesmo possui aproximadamente 25% das linhas com dados faltantes. Sendo necessário o preenchimento dessas lacunas para não descartar esses dados tão importantes. Para tanto, utilizou-se da imputação de dados única separadamente para quatro variáveis do banco de dados, nas quais tinham maior porcentagem e importância para o estudo da rodovia. Realizou-se a aplicação de modelos de previsão e regressão, totalizando o uso de 15 modelos comparando-os e escolhendo os de melhores resultados, onde foram aplicados para a imputação de dados. Os modelos empregados foram: Regressão linear, Regressão de Vetores de Suporte (RVS), Rede Neural Artificial (RNA), Árvore de decisão, Floresta Aleatória, *XGBoost* e os modelos otimizados para Árvore de Decisão, Floresta Aleatória, *XGBoost* com três técnicas de otimização. Após aplicado, o método foi comparado às técnicas de imputações de dados mais utilizadas atualmente, onde obteve resultados favoráveis para com a imputação de valores de dados faltantes possuindo precisão superior para a maioria das colunas propostas.

PALAVRAS-CHAVE: Dados geotécnicos; Imputação de dados; Dados faltantes; Rodovias.

ABSTRACT

The prediction of geotechnical data has been essential for the execution of road projects and highway maintenance, reducing costs and improving quality. By incorporating statistical techniques, it is possible to effectively predict pavement conditions and soil tests. Thus, the objective of this study is to impute data correlated to the fundamental variables in the database derived from 47 road projects of the National Department of Transport Infrastructure (DNIT) and 10 projects from the Ceará Public Works Department (SOP/CE), where approximately 25% of the lines have missing data. Filling these gaps is necessary to avoid discarding such important data. To this end, single data imputation was used separately for four variables in the database, which had the highest percentages and importance for the highway study. Prediction and regression models were applied, totaling the use of 15 models, comparing them and selecting the ones with the best results, which were then applied for data imputation. The models used were: Linear Regression, Support Vector Regression (SVR), Artificial Neural Network (ANN), Decision Tree, Random Forest, *XGBoost*, and optimized models for Decision Tree, Random Forest, and *XGBoost* with three optimization techniques. After application, the method was compared to the most currently used data imputation techniques, obtaining favorable results for imputing missing data values, with superior accuracy for most of the proposed columns.

KEYWORDS: Geotechnical data; Data imputation; Missing data; Highways.

¹ Universidade Federal do Cariri, e-mail: dvsdiegovieira@gmail.com; junior.peixoto@aluno.ufca.edu.br

² Universidade Federal do Ceará, e-mail: marigoncalves241@gmail.com; suelly@det.ufc.br

³ Instituto Federal de Educação, Ciência e Tecnologia do Ceará, e-mail: brunna.mikaelli@gmail.com; ajar.junior@gmail.com



INTRODUÇÃO

Com a finalidade de garantir a concepção eficaz de projetos rodoviários, é necessário que se tenha conhecimento acerca das diversas variáveis físicas provenientes do meio no qual serão executados, principalmente características relacionadas à topografia e a geotecnia (GUILHERME et al., 2020). Com intuito de obter essas informações com maior praticidade e com o advento das técnicas de ciência dos dados e aprendizado de máquina tem-se percebido uma crescente utilização dessas metodologias para a previsão de dados geotécnicos de rodovias, podendo se tornar uma prática importante para evitar o retrabalho e desperdício de recursos já empregados em projeto, garantindo assim qualidade e economicidade aos futuros projetos rodoviários.

Utilizando tais técnicas, é possível prever uma variedade de parâmetros geotécnicos, incluindo a densidade do solo, a umidade ótima e o índice de suporte califórnia (ISC). Essas predições são realizadas através de informações providas de ensaios básicos, como a granulometria, limite de liquidez, limite de plasticidade e compactação, de modo a estimar parâmetros mais complexos (SOUZA, RIBEIRO & SILVA, 2021). Essas estimativas permitem que os profissionais identifiquem áreas de potencial instabilidade, projetem estruturas de pavimentação adequadas e implementem medidas preventivas para mitigar os riscos de deslizamentos de terra, subsidência e outros problemas geotécnicos.

Ribeiro Silva e Barroso (2018) acrescentam que, o mapeamento de informações geotécnicas confiáveis podem ajudar a reduzir custos e aumentar a qualidade na elaboração e execução de projetos viários. Além disso, é possível preencher lacunas significativas, reduzindo a necessidade de novos ensaios caros e demorados, no caso de trechos rodoviários que não contenham informações. Ao extrapolar dados limitados de ensaios existentes para áreas semelhantes, esses métodos estatísticos podem fornecer estimativas confiáveis para uma variedade de parâmetros geotécnicos. Esse enfoque não apenas reduz os custos associados à realização de novos ensaios, mas também oferece uma abordagem eficaz para médias e pequenas obras rodoviárias, onde os recursos podem ser mais limitados.

De acordo com Soncim e Junior (2012), os modelos de previsão de dados geotécnicos integram um Sistema de Gerenciamento de Pavimentos (SGP), sendo empregados tanto em escala de rede quanto no planejamento, estimativa das necessidades globais de manutenção e reabilitação. Soncim, Junior e Campos (2014), explicam que, a função de um SGP é melhorar a eficiência das decisões e aumentar suas possibilidades, avaliar as consequências das decisões tomadas, facilitar a coordenação de atividades dentro de um órgão rodoviário. Se tornando assim, informações indispensáveis para o auxílio dos profissionais da área.

Na atualidade, não há como negar o crescimento exponencial dos dados disponíveis sobre os mais diversos assuntos permeando todos os aspectos das nossas vidas. Porém, mesmo tendo uma quantidade tão grande de informações, ainda enfrentamos um desafio recorrente em alguns conjuntos de dados: a presença de dados faltantes. Essa lacuna de dados (*missing data*) pode ocorrer pelos mais diversos motivos e afetar a proposta do conjunto de dados (*dataset*), como apontado por Ribeiro (2015). Esta situação possibilita a realização de estudos sobre a imputação de dados, que é uma área



amplamente abordada, visto sua colaboração em pesquisas e trabalhos acadêmicos, como apontado por Jäger, Allhorn e Bießmann (2021), o que nos leva ao estudo atual.

A imputação de dados faltantes se apresenta como uma ferramenta importante para o SGP, pois por meio dela será possível realizar uma análise dos dados mais precisa, ajudando a garantir a precisão, integridade e utilidade dos conjuntos de dados para tomada de decisões. Schreiber et al (2021), explica que a imputação de dados emergiu como um campo de pesquisa em expansão, impulsionado pela crescente instabilidade dos conjuntos de dados atuais. Esse fenômeno é resultado do frenético aumento na quantidade de dados, chamados também de big data, gerados ao longo dos anos, impulsionado pelo avanço contínuo das tecnologias.

Diante disso, este trabalho teve como objetivo de imputar dados faltantes de variáveis no banco de dados restrito ao território do estado do Ceará, criado para o Departamento Nacional de Infraestrutura de Transportes (DNIT) com de dados oriundos de 47 projetos rodoviários fornecido pelo próprio DNIT e 10 projetos fornecidos pela Secretaria de Obras Públicas do Ceará (SOP/CE). O banco de dados foi produzido através da digitalização de dados de antiga tabelas dos relatórios de projetos rodoviários.

Para a imputação dos dados ausente, foi desenvolvido um modelo de previsão dos dados geotécnicos de sub-leito das rodovias do Estado do Ceará, podendo assim, completar as lacunas do banco de dados existente, prevendo os dados faltantes de umidade ótima (OT), densidade seca (DEN), Índice de Suporte Califórnia (ISC) e expansão (EXP).

METODOLOGIA

Para a concepção deste estudo, empregaram-se uma variedade de técnicas avançadas de ciência de dados com o intuito de manipular e processar os dados brutos do conjunto inicial. Posteriormente, uma análise minuciosa envolvendo a utilização de quinze modelos distintos de classificação foi conduzida, visando antever os padrões e comportamentos subjacentes aos dados analisados.

O *dataset* foi georreferenciado, como apresentado na Figura 01, destacando a ampla diversidade de amostras que foram coletadas. Primeiramente todas as 7 mil linhas de dados no *dataset* do DNIT foram expostas as análises descritivas, visando o melhor entendimento dos dados, assim no *dataset* foi separado as variáveis que seriam úteis para a previsão, deste modo enxugando o mesmo e utilizando as seguintes colunas:

- Estaca: Esta variável representa as marcações ao longo da extensão da rodovia, permitindo a identificação precisa de cada ponto de medição. Cada estaca tem um espaçamento de 20 metros de distância uma da outra.
- OT (%) - Umidade Ótima: Refere-se à porcentagem de umidade ótima para a compactação do solo. A umidade ótima é determinada através de ensaios de compactação, sendo crucial para garantir a estabilidade e durabilidade do pavimento.
- DEN (g/cm³) - Densidade Seca: Indica a massa específica seca do solo, fornecendo informações sobre sua compactação e resistência.



- ISC/CBR (%) - Índice de Suporte Califórnia (ISC) / Índice de Resistência Califórnia (CBR): Este índice quantifica a resistência do solo, sendo crucial na determinação da capacidade de suportar cargas e tráfego.
- EXP (%) - Expansão: Refere-se à porcentagem de expansão do solo, que pode impactar significativamente a estabilidade e integridade do pavimento.
- Granulometria (50,8 (#2") - 0,074 (#200)): Especifica a distribuição granulométrica do solo, ou seja, a proporção de diferentes tamanhos de partículas. A granulometria é determinada através de ensaios de peneiramento, onde se calcula a porcentagem de cada fração granulométrica em relação ao total de material pré-definido.
- Local/Rodovia: Indica a localização específica da rodovia, fornecendo contexto geográfico e ambiental para os dados coletados.

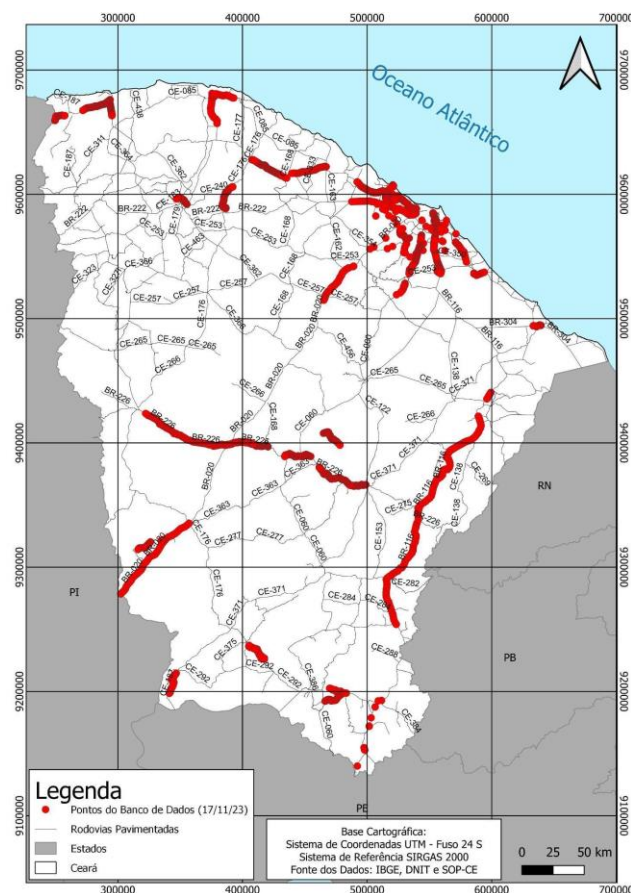


Figura 01. Mapa de localização dos pontos georreferenciados (Autores, 2024).

Tendo apenas Local/Rodovia como variável qualitativa, após separada as variáveis, foram feitos os pré-processamentos necessários, tais quais a tipagem correta dos dados e eliminação de dados errôneos. Oliveira (2022) corrobora em sua tese que, o pré-tratamento se faz necessário para provocar melhorias na interpretação e simplificar o modelo, fazendo com que seja mais robusto e confiável contra essas variações indesejáveis.



Para que fosse possível realizar a produção dos modelos de previsão de forma correta foi eliminado por hora todos os dados faltantes do *dataset*, após isso foram feitos os pré-processamentos comuns a todo estudo de previsão tais quais: *One-Hot Encoding* e a Padronização e Logaritmo, como apresentado na Figura 02.

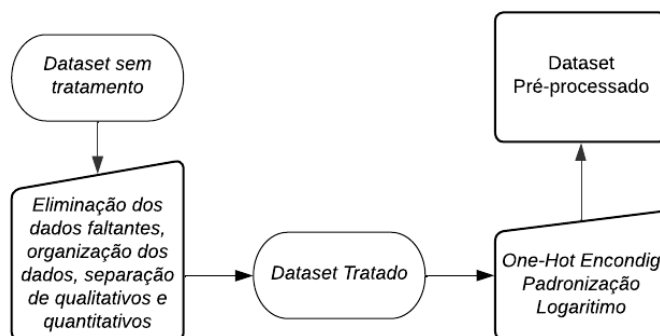


Figura 02. Fluxograma de Pré-Processamento (Autores, 2024).

Após o pré-processamento, foi feita a criação dos modelos. É importante ressaltar que para cada grupo de modelos foi retirada a variável em questão para produzir o modelo corretamente. No total, foram testados 15 modelos diferentes, sendo eles: Regressão Linear, Suporte à Regressão Vetorial (SRV), Rede Neural Artificial (RNA), Árvore de Decisão, Floresta Aleatória, *XGBoost* e os modelos otimizados para Árvore de Decisão, Floresta Aleatória, *XGBoost* com as otimizações mais usadas, sendo elas: *GridSearch*, *RandomizedSearch* e *BayesSearch*. Sendo o Floresta Aleatória e *XGBoost* os modelos com os melhores resultados, tais quais apresentaram respostas satisfatórias em outros trabalhos, como é explanado por Gioia, Barros e Silva (2022).

Para o treinamento dos modelos, os dados foram separados de acordo com os testes realizados. Sendo a separação de 85% dos dados para treino e 15% para teste, buscando evitar o *overfitting* e *underfitting*. Os modelos foram treinados sem otimização com parâmetros padrões e os modelos otimizados tiveram seus parâmetros calculados pelos algoritmos de otimização, assim sendo escolhido os melhores, com base na limitação da máquina a qual foi utilizada para realizar o estudo. Sendo ela um notebook Acer Nitro 5 com seguintes especificações: placa gráfica GeForce® GTX™ 1650 com memória dedicada VRAM de 4GB GDDR5, Processador Intel® Core™ i5-9300H Quad Core 9ª geração, além dos 16 GB de memória tipo DDR4. A Figura 03 ilustra o funcionamento do treinamento.

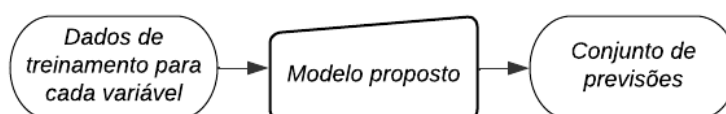


Figura 03. Fluxograma de treinamento dos modelos (Autores, 2024).

Depois disso foi realizada a imputação de dados utilizando os modelos treinados, no estudo em questão, em decorrência dos diversos dados faltantes, das quatro variáveis em diferentes linhas do *dataset*, cerca de 25% das linhas havia pelo menos uma das quatro colunas faltantes, então foi



proposta a seguinte solução: primeiramente as outras colunas que não as quatro focais do estudo foram inseridas por meio de imputação de dados simples, por meio do K-Vizinhos Próximos (*KNNImputer*) devido a sua baixa taxa de dados faltantes, estando todos por volta de 1%. Supardianto, Lalu Mutawali e Wafiah Murniati (2022) explicam que o *KNNImputer* é utilizado para lidar com conjuntos de dados que contêm valores ausentes, baseado na ideia de que observações semelhantes tendem a ter valores similares.

Feito isso foi selecionado o melhor modelo para cada variável e separadamente foi realizada a disjunção da coluna do *dataset*. Agora, com os dados faltantes, e devido a isso, muitas das linhas possuem mais de uma das colunas focais sem dados, faltando tanto para a variável separada quanto alguma outra das variáveis estudadas, logo para que os modelos preparados possam ser usados, têm que utilizar de algum método de imputação de dados comum, tal como o utilizado para os outros dados do *dataset*, sendo ele o *KNNImputer*, visto que o mesmo possui bons resultados no *dataset* em questão, assim, apenas no momento de utilização de cada modelo separadamente o banco esteja totalmente populado, fora a variável a qual é direcionada o modelo, só então é realizada a imputação de dados com o melhor modelo para cada coluna.

A imputação de dados por meio de modelos de previsão apresenta o seu melhor caso quando a coluna em questão é o único dado faltante entre os quatro da linha, e em pior caso temos as linhas onde todos as quatro colunas estão sem valor, porém devido ao as outras colunas do *dataset*, as quais foram utilizadas para o treinamento dos modelos e também aos valores imputados pelo *KNN Imputer*, é possível obter resultados satisfatórios até mesmo em piores casos.

Para que fosse possível validar a eficácia da solução utilizando os modelos de previsão foi desenvolvido um algoritmo capaz de verificar as métricas de erro dos valores imputados, primeiramente foi retirado todos os dados nulos do *dataset*, para que sempre houvesse como comparar o resultado da imputação de dados, então logo após foram retirados de forma pseudo aleatória 25% dos dados para cada variável focal, e só então os dados foram imputados e comparados aos dados reais, além de aplicada a técnica de imputação de dados envolvendo os modelos, também foram utilizados: *Simple imputer* utilizando a média como parâmetro descritivo e o *KNN imputer*, dois dos principais imputadores utilizados atualmente, os resultados obtidos foram favoráveis a técnica proposta neste estudo.

RESULTADOS E DISCUSSÃO

Após o processo de estudo do *dataset*, preparação dos dados, produção dos modelos e imputação de dados, foram obtidos dados relevantes para literatura e até mesmo para uso prático e profissional. Nesta seção serão expostos os resultados obtidos após a utilização do método proposto no *dataset* fornecido.

Primeiramente pode-se expor aqui os dados encontrados por meio da análise descritiva, onde foi retirado informações interessantes para a modelagem e análise de desempenho dos modelos, é possível observar os principais dados das variáveis na Tabela 1:



Tabela 1. Análises descritivas dos dados (Autores, 2024).

	ISC/CBR	OT	DEN	EXP
Média	18.995	10.088	1.920	0.164
Mediana	13.000	9.800	1.900	0.080
Moda	10.000	9.000	1.900	0
Desvio padrão	17.615	2.475	0.382	0.358
Amplitude	146.000	51.400	14.330	9.300
Variância	310.303	6.126	0.145	0.128

Após a análise descritiva tem-se os resultados obtidos para cada modelo feito e como exposto na metodologia foram utilizados 15 modelos, sendo 9 deles otimizados, não coincidentemente, os mesmos tiveram os melhores resultados para todas as variáveis. Sendo utilizada as seguintes métricas para avaliação do modelo:

- R^2 Score: é um coeficiente de determinação, isso implica que o R^2 está sendo usado como uma medida para avaliar a qualidade do ajuste do modelo (DUENHA et al., 2020). Quanto mais próximo de 1 o valor de R^2 indica um bom ajuste do modelo aos dados, enquanto um valor baixo indica um ajuste insatisfatório (Z. LI, S. LI e N. BIRBILIS, 2024). Os mesmos apresentam a Equação 1 para se obter o valor de R^2 Score:

$$R^2 \text{ Score} = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (1)$$

Onde:

N é o número total de dados;

y_i é o valor real

$\bar{y}$ é o valor médio de todos os valores observados;

$\hat{y}_i$ é o valor previsto.

- Erro médio Absoluto (MAE): quantifica a magnitude média dos erros de previsão em relação aos valores reais, sem considerar se os erros são positivos ou negativos (JONER, 2020). O seu valor é obtido de acordo com a Equação 2, apresentada pelo o mesmo autor:

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

Onde:

n é o número total de dados

y_i é o valor real

$\hat{y}_i$ é o valor previsto



- Erro Quadrático Médio da Raiz (RMSE): é uma métrica utilizada para avaliar a precisão de modelos de previsão ou estimativa em relação aos valores reais, quanto menor o valor do RMSE, melhor o desempenho do modelo, pois indica que as previsões estão, em média, mais próximas dos valores reais (HALLAK; PEREIRA FILHO, 2011). A Equação 3 é utilizada para calcular o RMSE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

Onde:

n é o número total de dados

y_i é o valor real

$\hat{y}_i$ é o valor previsto

- Erro Quadrático Médio da Raiz Logarítmica (RMSLE): ela irá avaliar a precisão de modelos de previsão, para fazer o seu cálculo, é adicionado 1 a ambos os valores reais e previstos antes de calcular o logaritmo natural para evitar problemas com o logaritmo de zero, permitindo a sua aplicabilidade mesmo que um dos valores seja zero, porém ambos valores não podem ser negativos (OLIVEIRA, 2019). A Equação 4 mostra como o seu valor pode ser calculado:

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(y_i + 1) - \log(\hat{y}_i + 1))^2} \quad (4)$$

Onde:

n é o número total de dados

y_i é o valor real

$\hat{y}_i$ é o valor previsto

Munido dos dados pré-processados, e a separação de teste do *dataset*, os modelos foram testados com 15% dos dados totais como proposto pela metodologia, e então foram obtidos os resultados em forma das métricas expostas. Para melhor entendimento, os resultados dos 3 melhores modelos serão demonstrados a seguir por meio de Tabelas 2, 3, 4 e 5 separadas por variável:

Tabela 2. Desempenho dos modelos para variável ISC/CBR (AUTORES, 2024).

ÍNDICE DE SUPORTE CALIFÓRNIA (ISC) / ÍNDICE DE RESISTÊNCIA CALIFÓRNIA (CBR)

Modelo	R ² Score	MAE (%)	RMSE (%)	RMSLE
Floresta aleatória <i>Bayes Search</i>	0,814	0,251	0,469	0,117
Floresta aleatória <i>Grid Search</i>	0,813	0,248	0,471	0,117



Floresta aleatória <i>Randomized Search</i>	0,811	0,254	0,473	0,117
--	-------	-------	-------	-------

Tabela 3. Desempenho dos modelos para variável Umidade Ótima (AUTORES, 2024).

UMIDADE ÓTIMA (OT)				
Modelo	R ² Score	MAE (%)	RMSE (%)	RMSLE
Floresta aleatória <i>Grid Search</i>	0,651	0,399	0,571	0,164
Floresta aleatória <i>Randomized Search</i>	0,649	0,399	0,573	0,151
Floresta aleatória <i>Bayes Search</i>	0,644	0,41	0,576	0,163

Tabela 4. Desempenho dos modelos para variável Densidade Seca (AUTORES, 2024).

DENSIDADE SECA (DEN)				
Modelo	R ² Score	MAE (g/cm ³)	RMSE (g/cm ³)	RMSLE
Floresta aleatória <i>Bayes Search</i>	0,710	0,360	0,522	0,135
Floresta aleatória <i>Grid Search</i>	0,708	0,361	0,524	0,179
SRV	0,695	0,397	0,535	0,169

Tabela 5. Desempenho dos modelos para variável Expansão (AUTORES, 2024).

EXPANSÃO (EXP)				
Modelo	R ² Score	MAE (%)	RMSE (%)	RMSLE
<i>XGBoost Bayes Search</i>	0,327	0,334	1,08	0,191
Floresta aleatória <i>Grid Search</i>	0,325	0,322	1,082	0,187
Floresta aleatória <i>Bayes Search</i>	0,312	0,330	1,092	0,190

Verificados resultados razoavelmente satisfatórios para principalmente para o R² Score na maioria das variáveis, onde quanto mais próximo de 1 melhor explica a variabilidade dos dados, vemos também que para todos os modelos o RMSE é maior que o MAE levando a entender que existem *outliers* nos dados, Aires (2022) explica que essa comparação pode servir como um indicativo da presença de *outliers* nos dados, o que é corroborado com a análise descritiva, esses *outliers* podem afetar negativamente os modelos de previsão, porém os resultados obtidos sobre o Erro Médio Absoluto e as outras métricas, se faz perceber que comparado à análise descritiva os modelos apresentam um bom resultado. Logo, para próxima etapa foram utilizados os modelos com melhores



métricas dentre os testados, sendo eles: Floresta aleatória *Bayes Search*, Floresta aleatória *Grid Search*, Floresta aleatória *Bayes Search* e *XGBoost Bayes Search*, respectivamente para as variáveis: ISC/CBR, OT, DEN, EXP.

Passando para fase de imputação de dados, foram obtidos ótimos resultados quando comparados a outros métodos de imputação mais utilizados atualmente, se utilizando da proposta de validar a técnica idealizada pelo estudo, foi realizado o teste para a imputação dos dados separando os dados não nulos e aplicando a remoção de 25% dos dados para cada variável focal e logo após a imputação de dados e validação como proposto pela metodologia, os resultados obtidos foram os seguintes:

Tabela 6. Resultados Imputador Simples Media (Autores, 2024).

	R² Score	MAE	RSME
OT (%)	0.711	0.192	0.533
DEN (g/cm ³)	0.735	0.067	0.176
ISC/CBR (%)	0.764	0.154	0.476
EXP (%)	0.655	0.133	0.584

Tabela 7. Resultados KNN Imputer (Autores, 2024).

	R² Score	MAE	RSME
OT (%)	0.824	0.132	0.416
DEN (g/cm ³)	0.905	0.038	0.105
ISC/CBR (%)	0.913	0.085	0.288
EXP (%)	0.835	0.095	0.403

Tabela 8. Resultados da metodologia proposta (Autores, 2024).

	R² Score	MAE	RSME
OT (%)	0.941	0.086	0.239
DEN (g/cm ³)	0.283	0.112	0.289
ISC/CBR (%)	0.894	0.073	0.319
EXP (%)	0.940	0.074	0.243

Com os resultados expostos pode-se perceber que para todas as variáveis exceto o DEN (g/cm³), possivelmente devido a escala de seus dados e sua pequena variância, temos resultados comparáveis e até superiores para o R² Score, e para o Erro médio absoluto tem-se resultados melhores que os dois imputadores utilizados, entendendo-se que o método é válido e oferece resultados ótimos quando comparado a métodos de imputação de dados já conhecidos.

CONCLUSÃO

Através desta implementação, a técnica foi capaz de imputar valores de dados faltantes com precisão para a maioria das colunas propostas, um ato que mantém não apenas a integridade e a qualidade do conjunto de dados, mas também serve para futuros estudos dependentes de dados. Além disso, a adoção do método proposto garante a eliminação de investimentos em testes de campo redundante



onde a imputação de dados pode ser realizada rapidamente. Isto leva a uma conservação substancial dos gastos financeiros e temporais.

O sucesso da metodologia proposta pode ser apreciado pelas métricas reveladas através dos resultados. Quando vista juntamente com os métodos de imputação de dados convencionais, a comparação sublinha a sua competitividade e viabilidade, indicando assim que tem-se uma forte razão para considerar esta abordagem de forma favorável. Com esta informação, é seguro dizer que a solução aqui apresentada é eficaz e promissora para abordar valores faltantes, uma abordagem que deve formar uma base sólida sobre quais análises e tomadas de decisão adicionais podem ser realizadas em contextos profissionais ou acadêmicos.

REFERÊNCIAS

- AIRES, U. R. V. Modelagem temporal e espacial da concentração superficial de sedimentos utilizando sensoriamento remoto orbital e aprendizado de máquina. 2022. 114 f. Tese (Doutorado em Engenharia Agrícola) - Universidade Federal de Viçosa, Viçosa. 2022.
- DUENHA, L.; SONOHATA, R.; ARIGONI, D.; SANTOS, R. Preditor de Desempenho de GPUs aplicado à Exploração do Espaço de Projetos ciente de Dark Silicon. In: SIMPÓSIO EM SISTEMAS COMPUTACIONAIS DE ALTO DESEMPENHO (SSCAD), 21. , 2020, Online. Porto Alegre: Sociedade Brasileira de Computação, 2020 . p. 299-310.
- Gioia, T.; Barros, J. R.; Silva, R. R. (2022). Fatores socioeconômicos e algoritmos de machine learning aplicados à predição de risco de doenças negligenciadas.: Estudo de caso nos municípios do Estado de Goiás e Distrito Federal, Brasil. *Finisterra*, 57(121), 109–123.
- Guilherme, A. T. P., da Silva, C. A. U., Ribeiro, A. J. A., & de Araújo Barroso, S. H. (2020). Previsão estatística e mapeamento das propriedades geotécnicas dos solos da microrregião de mossoró-RN. *Brazilian Journal of Development*, 6(4), 22297-22318.
- HALLAK, R.; PEREIRA FILHO, A. J. Metodologia para análise de desempenho de simulações de sistemas convectivos na região metropolitana de São Paulo com o modelo ARPS. *Revista Brasileira de Meteorologia*, v. 26, n. 4, p. 591–608, 2011.
- JÄGER S.; ALLHORN A.; BIESSMANN F. (2021). Um benchmark para métodos de imputação de dados. *Frente. Grandes Dados* 4:693674. doi: 10.3389/fdata.2021.693674
- JONER, Henrique. Inferência preditiva geoespacial da criminalidade em Porto Alegre: uma abordagem de aprendizado de máquina. 2020. Dissertação Programa de Pós-Graduação em Economia da Faculdade de Ciências Econômicas da UFRGS, Porto Alegre - RS, 2020.
- LI, Z.; LI, S.; BIRBILIS, N. A machine learning-driven framework for the property prediction and generative design of multiple principal element alloys. *Materials Today Communications*, v. 38, p. 107940, 2024.
- OLIVEIRA, F. T. M.. Predição do tempo de viagem com base em trajetórias automobilísticas. 2019. 50 f. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação)-Universidade Federal do Ceará, Campus de Quixadá, Quixadá, 2019.
- OLIVEIRA, R. Co. G. Previsão de ozônio troposférico na região metropolitana do Rio de Janeiro com base em técnicas de imputação de dados faltantes e calibração multivariada. 2022. 144 f. Tese (Doutorado em Engenharia Ambiental) - Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2022.



RIBEIRO, A. J. A; SILVA, C. A. U; BARROSO, S. H. A. (2018). Metodologia de baixo custo para mapeamento geotécnico aplicado à pavimentação. *Transportes*, v.26, p.84-100.

RIBEIRO, E. A. Imputação de dados faltantes via algoritmo EM e rede neural MLP com o método de estimativa de máxima verossimilhança para aumentar a acurácia das estimativas. 2015. 162 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Sergipe, São Cristóvão, 2015.

SCHREIBER, Jonas F. et al. Técnicas de imputação de dados aplicadas ao ambiente dassmart grids: uma revisão. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, v. 8, n. 1, 2021.

SONCIM, S. P., & JÚNIOR, J. L. F. (2012). Desenvolvimento de um modelo de desempenho de irregularidade longitudinal para a rede de rodovias em tratamento superficial duplo do estado da Bahia. *Transportes*, 20(1), 13-21.

SONCIM, S. P., JÚNIOR, J. L. F., de CAMPOS, L. E. P. (2014). Modelo de previsão da área de desgaste em pavimentos rodoviários com revestimento em tratamento superficial. *Transportes*, 22(2), 51-61.

SOUZA, W. M. D., RIBEIRO, A. J. A., & SILVA, C. A. U. D. (2021). Uma abordagem alternativa para obtenção da classificação de solos da AASHTO usando redes neurais artificiais.

SUPARDIANTO, LALU Mutawalli, and WAFIAH Murniati, “Penerapan Knnimputer Dalam Mengolah Data Missing Value Untuk Membantu Meningkatkan Akurasi Support Vector Machine Klasifikasi Penyakit Tiroid”, *JINTEKS*, vol. 4, no. 4, pp. 386-390, Nov. 2022.

AGRADECIMENTOS

Os pesquisadores agradecem ao DNIT, a CAPES, a Universidade Federal do Cariri (UFCA), a Universidade Federal do Ceará (UFC) e ao Instituto Federal do Ceará (IFCE) pelo apoio e financiamento à pesquisa.