UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
CAMPUS DE CHAPADÃO DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

LEONARDO BEZERRA DA SILVA

PREDIÇÃO DE PRODUTIVIDADE DE SEMENTES DE SOJA USANDO ATRIBUTOS DO SOLO E APRENDIZAGEM DE MÁQUINA

CHAPADÃO DO SUL – MS
2023

UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
CAMPUS DE CHAPADÃO DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

LEONARDO BEZERRA DA SILVA

PREDIÇÃO DE PRODUTIVIDADE DE SEMENTES DE SOJA USANDO ATRIBUTOS DO SOLO E APRENDIZAGEM DE MÁQUINA

Orientador: Prof. Dr. Ricardo Gava

Dissertação apresentada à Universidade Federal de Mato Grosso do Sul, como parte dos requisitos para obtenção do título de Mestre em Agronomia, área de concentração: Produção Vegetal.

CHAPADÃO DO SUL – MS
2023

Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

CERTIFICADO DE APROVAÇÃO

DISCENTE: Leonardo Bezerra da Silva

ORIENTADOR: Dr. Ricardo Gava

TÍTULO: Predição de produtividade de sementes de soja usando atributos do solo e aprendizagem de máquina

AVALIADORES:

Prof. Dr. Ricardo Gava

Profa. Dra. Larissa Pereira Ribeiro Teodoro

Prof. Rafael Ferreira Barreto

Chapadão do Sul, 16. de junho de 2023.

Documento assinado eletronicamente por **Ricardo Gava**, **Professor do Magisterio Superior**, em 16/06/2023, às 14:27, conforme horário oficial de Mato Grosso do Sul, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Larissa Pereira Ribeiro**, **Professora do Magistério Superior**, em 16/06/2023, às 14:27, conforme horário oficial de Mato Grosso do Sul, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rafael Ferreira Barreto**, **Professor do Magisterio Superior**, em 16/06/2023, às 14:27, conforme horário oficial de Mato Grosso do Sul, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

“Eu acredito que as vezes são as pessoas que ninguém espera nada, que façam as coisas que ninguém consegue imaginar.”

Alan Turing

AGRADECIMENTOS

Ao orientador Prof. Dr. Ricardo Gava, por ter acreditado desde o começo nessa ideia, me orientando e apoiando ao longo destes anos

À João Airton Malta Feitosa Filho, que me incentivou, me dando suporte e estímulo para buscar meus objetivos.

À empresa ATTO Sementes, por todos esses anos de parceria e por fazer possível à obtenção de tais conhecimentos durante o meu período de trabalho, em especial, ao meu líder Elder Landgraf.

À todos os meus amigos, em especial à Edvar Nantes Lopes Junior, que não pode ver a conclusão deste projeto, porem sempre me apoiou com a sua animação, me inspirando a ser e fazer melhor pela agricultura que tanto amava, espero que esteja orgulhoso.

LISTA DE FIGURAS

Página

Figura 1..... 11

Figura 2..... 14

Figura 3..... 17

Figura 4..... 16

RESUMO

Existem diversas pesquisas que buscaram entender como os atributos químicos do solo influenciam a produtividade de grãos de soja. Essas pesquisas utilizaram, em sua maioria, técnicas de geoestatística e análise multivariada para demonstrar o efeito de variáveis como teor de matéria orgânica, CTC, argila sobre os componentes produtivos e desempenho fisiológico da soja. O objetivo da pesquisa foi testar diferentes atributos químicos como entrada em modelos de aprendizagem de máquina para estimar a produtividade de sementes de soja. O presente trabalho foi conduzido utilizando o banco de dados (BD) de solo e produtividade de grãos da sementeira ATTO Sementes para a cultura da soja cultivada na safra 2020/2021. Os atributos de solo avalidos foram pH, capacidade de troca catiônica (CTC), saturação por bases (V%), teor de argila, matéria orgânia (MO). A produtividade foi obtida por meio da geração de um mapa de colheita, pelo sistema integrado JDLink da Colhedora John Deere S790. A correlação de Pearson foi realizada para verificar a interrelação entre as variáveis de solo analisadas e a produtividade. Os dados foram submetidos as análises de aprendizagem de máquina (redes neurais artificiais, regressão linear, M5P, REPTree, random forest e máquina de vetor suporte). Foram testadas seis configurações de conjuntos de dados de entrada: pH, CTC, V%, altitude, argila e todas as informações juntas. Como variável de saída (output) dos algoritmos foi utilizado a produtividade da soja A utilização de todas as variáveis de solo (pH, CTC, SB, teor de argila e MO) associado ao modelo de aprendizagem de máquina random forest possibilita predizer a produtividade de sementes de soja com alta precisão.

Palavras-chave: *Glycine* max, pH, teor de argila, random forest.

ABSTRACT

There are several studies that sought to understand how the chemical attributes of the soil influence the yield of soybeans. These researches used, for the most part, geostatistical techniques and multivariate analysis to demonstrate the effect of variables such as organic matter content, CEC, clay on the productive components and physiological performance of soybean. The objective of the research was to test different chemical attributes as input in machine learning models to estimate soybean seed productivity. The present work was carried out using the database (BD) of soil and grain yield of the ATTO Sementes sowing for the soybean crop grown in the 2020/2021 season. The evaluated soil attributes were pH, cation exchange capacity (CEC), base saturation (SB), clay content, organic matter (OM). Seed yield was obtained through the generation of a harvest map, by the integrated JDLink system of the John Deere S790 Harvester. Pearson's correlation was performed to verify the interrelationship between the analyzed soil variables and productivity. The data were subjected to machine learning analysis (artificial neural networks, linear regression, M5P, REPTree, random forest and support vector machine). Six different configurations for the algorithms were tested: pH, CTC, V%, altitude, clay and all information together. As an output variable (output) of the algorithms, soy bean productivity was used. The use of all soil variables (pH, CEC, SB, clay content and MO) associated with the random forest machine learning model makes it possible to predict soybean seed yield with high precision.

Keywords: *Glycine* max, pH, clay content, random forest.

RESUMO

INTRODUÇÃO 10

MATERIAL E MÉTODOS 11

RESULTADOS E DISCUSSÃO 14

CONCLUSÕES 19

REFERENCIAS 19

INTRODUÇÃO

O sucesso da lavoura de soja depende de diversos fatores, mas sem dúvida, o mais importante deles é a utilização de sementes de elevada qualidade, que geram plantas de alto vigor, que terão um desempenho superior no campo. O uso de semente de elevada qualidade permite o acesso aos avanços genéticos, com as garantias de qualidade e tecnologias de adaptação nas diversas regiões, assegurando maiores produtividades. Portanto, o estabelecimento da lavoura de soja com sementes da mais alta qualidade é de fundamental importância. (EMBRAPA, 2016).

A qualidade da semente de soja pode ser influenciada por diversos fatores, que podem ocorrer durante a fase de produção no campo, na operação de colheita na secagem, no beneficiamento, no armazenamento, no transporte e na semeadura. Dentre esses fatores, Mattioni et al. (2013) destacam que atributos químicos do solo podem afetar a quantidade e a qualidade final das sementes de soja produzida. Mondo et al. (2012) encontraram correlação positiva e significativa entre matéria orgânica e germinação de sementes de soja. Contudo, Mattioni et al. (2013) não observaram correlação significativa entre essas variáveis.

Existem diversas pesquisas que buscaram entender como os atributos químicos do solo influenciam a produtividade de grãos de soja (VIDEIRA et al., 2019; LOVERA et al., 2018; DALCHIAVON et al., 2017; ROSA FILHO et al., 2009). Essas pesquisas utilizaram, em sua maioria, técnicas de geoestatística e análise multivariada para demonstrar o efeito de variáveis como teor de matéria orgânica, CTC e argila sobre os componentes produtivos e desempenho fisiológico da soja. Tais análises, apesar de importantes, não exploram de forma precisa a relação não-linear que pode ocorrer entre as variáveis.

Diante disso, diversas técnicas de aprendizagem de máquina podem ser utilizadas para estimar a produtividade de soja com alta precisão. Dentre as técnicas disponíveis destacam-se os algoritmos de floresta aleatória (*random forest*), redes neurais artificiais, árvores de decisão (M5P, REPTree), dentre outros. Teodoro et al. (2021) avaliaram diferentes algoritmos de aprendizagem de máquina para estimar a produtividade de grãos de soja utilizando variáveis espectrais coletadas com aeronave remotamente pilotada. Os autores demonstraram que o melhor algoritmo foi o de floresta aleatória. Santana et al. (2023) demonstraram que é possível classificar genótipos de soja quanto a características

industriais como teor de óleo e proteína usando esse mesmo algoritmo associado a variáveis espectrais.

Sendo assim, a hipótese a ser testada nesse trabalho foi de ser possível predizer a produtividade de sementes de soja usando atributos químicos do solo associados a modelos de aprendizagem de máquina. O objetivo da pesquisa foi testar diferentes atributos químicos como entrada em modelos de aprendizagem de máquina para estimar a produtividade de sementes de soja.

MATERIAL E MÉTODOS

Coleta dos dados

O presente trabalho foi conduzido utilizando o banco de dados (BD) de solo e produtividade de grãos da sementeira ATTO Sementes para a cultura da soja cultivada na safra 2020/2021. Esse BD foi gerado utilizando plataformas digitais para armazenamento e processamento dos dados de execução. Foi selecionado o talhão 3403 com 244,9 ha, localizado na Fazenda Ribeirão da Velhas (Latitude: -17.1115750 Longitude:-53.9603195) localizada no município Alto Garças, Mato Grosso.

Os atributos de solo avaliados foram pH, capacidade de troca catiônica (CTC), saturação por bases (SB), teor de argila, matéria orgânica (MO) foram coletados a cada 30 ha, sendo realizada a coleta de três amostras por ponto, totalizando 72 amostras finais. As médias dos valores dos 24 pontos estão contidas na Figura 1.

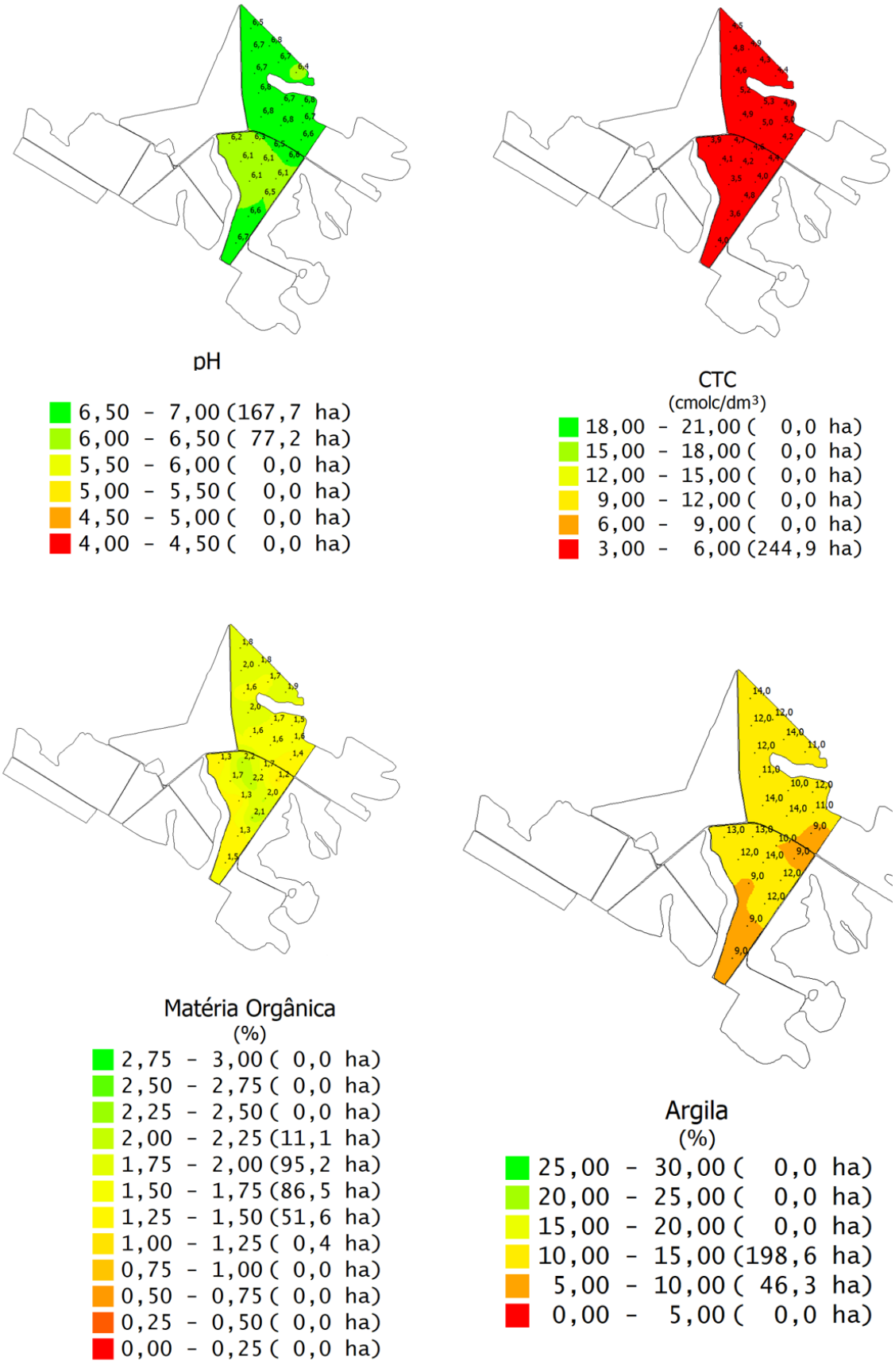

Figura 1. Mapa de distribuição dos atributos químicos de solo avaliados.

A produtividade foi obtida por meio da geração de um mapa de colheita, pelo sistema integrado JDLink da Colhedora John Deere S790. Após esse processo, houve uma sobreposição entre os mapas gerados pela amostragem de solo e o mapa de colheita, onde após criação de pixels de mesmo tamanho entre os mapas sobrepostos, foram obtidos todos os atributos, correlacionados a um mesmo posicionamento, proveniente de fontes diferentes.

Análises estatísticas e de inteligência computacional

A correlação de Pearson (Equação 1) foi realizada para verificar a interrelação entre as variáveis de solo analisadas e a produtividade. Foi utilizado um mapa de calor de correlação para expressar graficamente os resultados. Correlações positivas foram expressas em azul e correlações negativas foram expressas em vermelho, de acordo com a seguinte equação:

$$r(xy) = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}.\sqrt{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

Em que: x1 , x2 , ..., xn e y1 , y2 , ..., yn são os valores medidos de ambas as variáveis; $\bar{\boldsymbol{x}}$ e $\bar{\boldsymbol{y}}$ são as médias aritméticas das variáveis.

Posteriormente os dados foram submetidos as análises de aprendizagem de máquina (Tabela 1), testando seis configurações de entrada (input): pH, CTC, V%, altitude, argila e todas as informações juntas e a técnica de Regressão Linear Múltipla foi utilizada como modelo controle. Como variável de saída (output) dos algoritmos foi utilizada a produtividade de soja. A predição foi realizada por meio de validação cruzada estratificada com k-fold = 10 e dez repetições (100 execuções para cada modelo). Todos os parâmetros dos modelos foram estabelecidos de acordo com a configuração default do software Weka 3.8.5.

Tabela 1. Relação dos modelos de aprendizagem de máquina utilizados na predição da produtividade de soja.

Sigla	Modelo de aprendizagem de máquinas	Referência

ANN	Redes neurais artificiais	(Egmont-Petersen et al., 2002)
LR	Regressão Linear Múltipla	(Štepanovský et al., 2017)
M5P	Árvore de decisão M5P	(Quinlan, 1993)
REPTree	Árvore de decisão REPTree	(Snousy et al., 2011)
RF	Floresta aleatória	(Belgiu & Drăguţ, 2016)
SVM	Máquina de vetor suporte	(Nalepa & Kawulok, 2019)

Nas avaliações do desempenho dos modelos de predição testados foram usadas métricas de coeficiente de correlação de Pearson (r) e Erro Absoluto Médio (MAE). Para verificar a significância dos inputs e as técnicas de Machine Learning testados, e a interação entre ambos, foi realizado uma análise de variância. Havendo a presença da significância, foram gerados boxplots com as médias de r e MAE, agrupados pelo teste de Scott-Knott (Scott & Knott, 1974) ao nível de 5% de probabilidade. O agrupamento das médias foi realizado como o auxílio do software Rbio, enquanto os boxplots foram gerados usando os pacotes ggplot2 e ExpDes.pt do software R.

RESULTADOS E DISCUSSÃO

A Figura 2 contém a correlação de Pearson entre as variáveis avaliadas. As maiores correlações foram observadas de forma positiva entre pH e saturação por bases (V%) e capacidade de troca catiônica (CTC) e matéria orgânica. Resultados similares foram observados em outras pesquisas com a cultura da soja (VIDEIRA et al., 2019; LOVERA et al., 2018; DALCHIAVON et al., 2017; ROSA FILHO et al., 2009). É importante destacar que a produtividade de sementes (Prod) de soja não se correlacionou em alta magnitude com nenhum atributo de solo avaliado. Esses resultados demonstram que não existe associação linear entre os atributos e a Prod. Teodoro et al. (2021) relatam que nestas condições (ausência de associação linear), os algoritmos de aprendizagem de máquina podem ser utilizados para predição da variável principal (Prod), pois possuem a capacidade de encontrar padrões não-lineares entre as variáveis avaliadas.

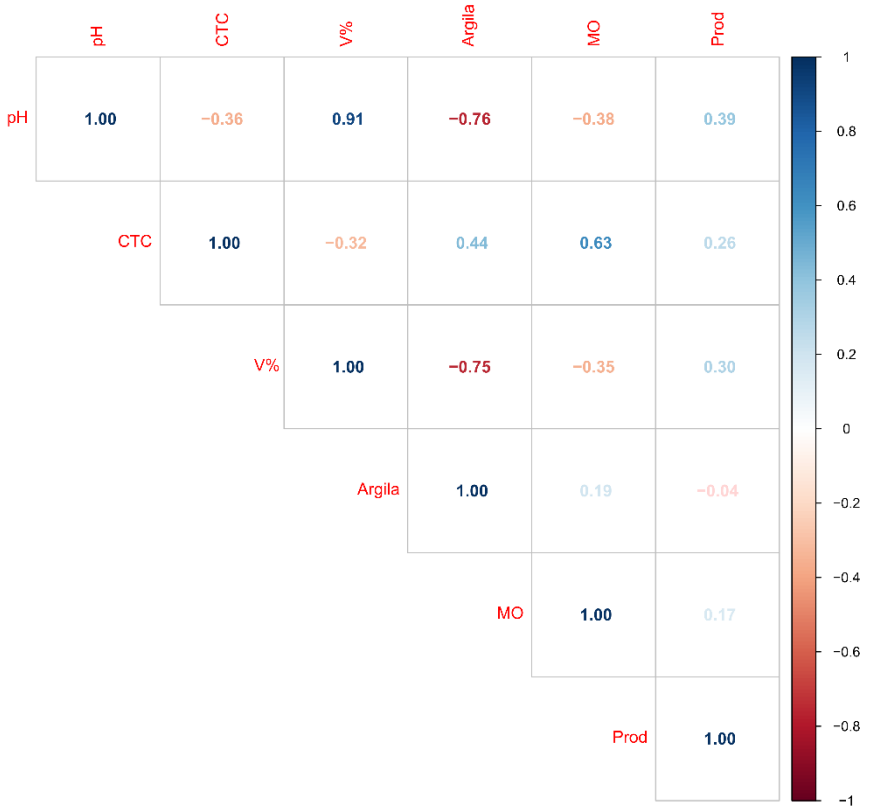

Figura 2. Mapa de calor para a análise de correlação de Pearson entre as variáveis de solo analisadas e a produtividade.

Correlações positivas expressas em azul e correlações negativas expressas em vermelho, intensidade de cores diz respeito a magnitude da correlação, quanto mais escura a cor mais forte é a correlação entre as variáveis.

A Tabela 2 contém os efeitos de input e algoritmos de aprendizagem de máquina testados quanto as estimativas do coeficiente de correlação (r) e para erro absoluto médio (MAE). Houve interação significativa entre input e algoritmos de aprendizagem de máquina para ambas as variáveis avaliadas. Esse resultado indica que o melhor algoritmo depende do input usado e vice-versa. Resultados similares foram observados por Teodoro et al. (2021), que observaram a interação entre diferentes inputs e algoritmos de aprendizagem de máquina para a predição da produtividade de grãos de soja.

Tabela 2. Resumo da análise de variância para o coeficiente de correlação (r) e para erro absoluto médio (MAE) para os diferentes inputs testados e algoritmos de aprendizagem de máquina (ML) testados.

F.V	GL	r	MAE
Input	5	3.47*	31.587*
Aprendizagem de máquina (ML)	5	0.9837*	6.3402*
Input	25	0.0357*	0.7226*
Resíduo	324	0.003	0.0016
CV (%)		1.42	1.67

F.V: fonte de variação; G.L: graus de liberdade; M.L: aprendizado de máquinas; CV: coeficiente de variação; r: coeficiente de correlação; MAE: erro absoluto médio.

A Figura 3 contém o agrupamento de médias para a interação significativa entre inputs e algoritmos de aprendizagem de máquina para o coeficiente de correlação de Pearson (r) entre os valores estimados e preditos. O algoritmo M5P proporcionou as maiores médias de r quando são usados apenas um dos atributos de solo como input de forma indiscriminada quanto a qual atributo. Contudo, quando todos os atributos são utilizados juntos, o algoritmo random forest apresentou a maior média. Esses resultados corroboram os obtidos por Teodoro et al. (2021) e Santana et al. (2023), que encontraram superioridade do random forest em relação aos demais algoritmos testados. Ao comparar os diferentes inputs, observa-se que a utilização de todos os atributos de solo simultaneamente proporciona os maiores valores de r, independente do algoritmo de aprendizagem de máquina. Além disso, é importante ressaltar a inferioridade demonstrada pelo modelo de Regressão Linear Múltipla com relação a todos os modelos de aprendizagem, que mesmo utilizando todas as variáveis, o modelo tradicional apresentou a pior performance, juntamente com SVM.

Figura 3. Boxplot para coeficiente de correlação (r) para o agrupamento de médias entre os inputs (pH, CTC, V%, altitude, argila e todas as informações) dos algoritmos (ANN, LR, M5P, REPTree, RF e SVM) na predição de produtividade de grãos de soja. Letras maiúsculas comparam as médias dos algoritmos de aprendizagem de máquina, enquanto letras minúsculas comparam os inputs pelo teste de Scott-Knott a 5% de probabilidade.

A comparação de médias para a interação significativa entre inputs e algoritmos de aprendizagem de máquina para erro médio absoluto (MAE) entre os valores estimados e preditos está contido na Figura 4. As menores médias de MAE com a utilização de apenas um dos atributos de solo como input foram obtidas com M5P. Contudo, quando todos os atributos são utilizados juntos, o algoritmo random forest apresentou a menor média. Ao comparar os diferentes inputs, verifica-se que a utilização de todas os atributos de solo simultaneamente proporciona os menores valores de MAE, independente do algoritmo de aprendizagem de máquina.

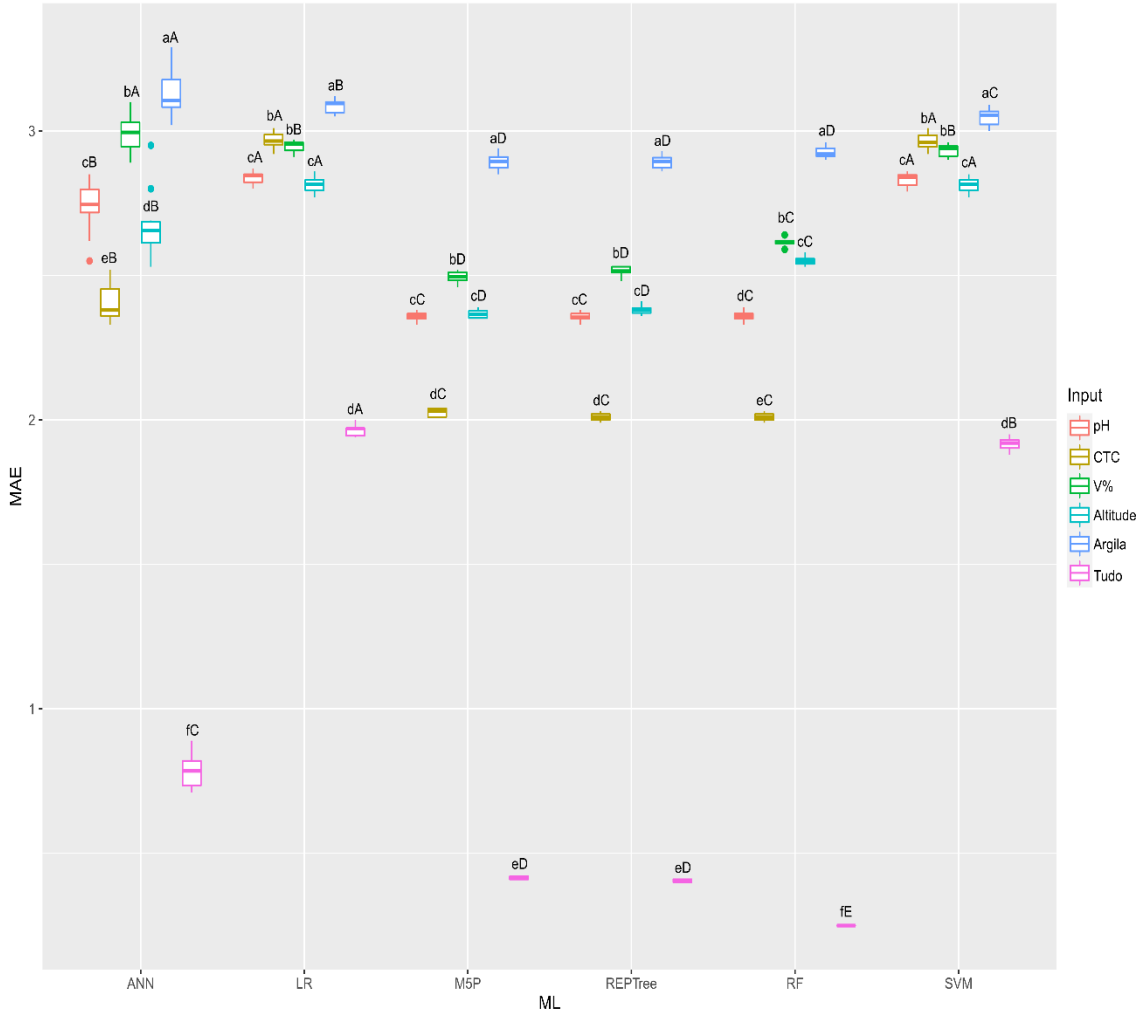

Figura 4. Boxplot para coeficiente de correlação (r) para o agrupamento de médias entre os inputs (pH, CTC, V%, altitude, argila e todas as informações) dos algoritmos (ANN, LR, M5P, REPTree, RF e SVM) na predição de produtividade de grãos de soja. Letras maiúsculas comparam as médias dos algoritmos de aprendizagem de máquina, enquanto letras minúsculas comparam os inputs pelo teste de Scott-Knott a 5% de probabilidade.

O algoritmo random forest associado a inclusão de todos os atributos de solo proporcionou a maior média de r e a menor média de MAE. Portanto, essa é a melhor combinação possível para a predição da produtividade de grãos de soja.

O presente trabalho apresentou resultados consideráveis quanto a possibilidade de predição de produtividade, devido a quantidade e qualidade de dados presentes no BD, que permitiu dados completos quanto a atributos de solo, englobando toda a área do talhão estudado, permitindo uma análise na escala de pixel, com assertividade considerável.

Para futuros trabalhos no segmento de predição de produtividade, a utilização de atributos de solo com variáveis de clima, retenção de água no solo ao longo da safra, presença e ausência de pragas, doenças e plantas daninhas, teto produtivo de cultivares, pode ampliar a capacidade e assertividade preditiva da ferramenta, a fim de se tornar um poderoso artificio para tomada de decisão em pequenas e grandes propriedades

CONCLUSÕES

A utilização de todas as variáveis de solo (pH, CTC, SB, teor de argila e MO) associado ao modelo de aprendizagem de máquina random forest possibilita predizer a produtividade de sementes de soja com alta precisão.

REFERENCIAS

Belgiu, M. & Dragu, T. L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **114**, 24–Egmont-Petersen, M., Ridder, D. & Handels, H. Image processing with neural networks a review. *Pattern Recognit.* **35**, 2279–2301. https://doi.org/10.1016/S0031-3203(01)00178-9 (2002).

Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Dalchiavon, F. C., Rodrigues, A. R., De Lima, E. S., Lovera, L. H., & Montanari, R. (2017). Variabilidade espacial de atributos químicos do solo cultivado com soja sob plantio direto. Revista de Ciências Agroveterinárias, 16(2), 144-154.

Lovera, L. H., Lima, E. D. S., Montanari, R., de Souza, Z. M., Farhate, C. V. V., Campos, M. C. C., & Torres, J. L. R. (2018). Geostatistical techniques applied to spatial distribution of macroorganisms in soybean crop. Australian Journal of Crop Science, 12(3), 357-364.

Mattioni, N. M., Schuch, L. O., Villela, F. A., Zen, H. D., & Mertz, L. M. (2013). Fertilidade do solo na qualidade fisiológica de sementes de soja. Revista Brasileira de Ciências Agrárias, 8(4), 656-661.

Mondo, V. H. V., Gomes Junior, F. G., Pinto, T. L. F., Marchi, J. L. D., Motomiya, A. V. D. A., Molin, J. P., & Cicero, S. M. (2012). Spatial variability of soil fertility and its

relationship with seed physiological potential in a soybean production area. Revista Brasileira de Sementes, 34, 193-201.

Müller, D. H., Camili, E. C., Scaramuzza, W. L. M. P., & Albuquerque, M. C. D. F. (2018). Variabilidade espacial da qualidade de sementes de soja e dos atributos químicos do solo em campo de produção no Cerrado. Journal of Seed Science, 40, 25-35.

Quinlan, R. J. Learning with Continuous Classes. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore, 343-348, 1992.

Rosa Filho, G., Carvalho, M. D. P., Andreotti, M., Montanari, R., Binotti, F. F. D. S., & Gioia, M. T. (2009). Variabilidade da produtividade da soja em função de atributos físicos de um Latossolo Vermelho distroférrico sob plantio direto. Revista Brasileira de Ciência do Solo, 33, 283-293.

Santana, D. C., Teodoro, L. P. R., Baio, F. H. R., dos Santos, R. G., Coradi, P. C., Biduski, B., ... & Shiratsuchi, L. S. (2023). Classification of soybean genotypes for industrial traits using UAV multispectral imagery and machine learning. Remote Sensing Applications: Society and Environment, 100919.

Snousy, M. B. A., El-Deeb, H. M., Badran, K. & Khlil, I. A. A. Suite of decision tree-based classification algorithms on cancer gene expression data. *Egypt. Inf. J.* **12**, 73–

Teodoro, P. E., Teodoro, L. P. R., Baio, F. H. R., da Silva Junior, C. A., dos Santos, R. G., Ramos, A. P. M., ... & Shiratsuchi, L. S. (2021). Predicting days to maturity, plant height, and grain yield in soybean: A machine and deep learning approach using multispectral data. Remote Sensing, 13(22), 4632.

Videira, L. M. L., Silva, P. R. T., Pereira, D. D. S., Montanari, R., Panosso, A. R., & Oliveira, C. F. (2019). Effect of farming systems on the spatial variability of soil physical properties and soybean yield. Journal of Agricultural Science (Toronto), 11(15), 87-96.