

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
SETOR DE CIÊNCIAS AGRÁRIAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

PABLO HENRIQUE FERREIRA

REDUÇÃO DA DIMENSIONALIDADE PARA ESTIMATIVA DE TEORES DE
NUTRIENTES EM FOLHAS E GRÃOS DE SOJA COM ESPECTROSCOPIA NO
INFRAVERMELHO

PONTA GROSSA
2017

PABLO HENRIQUE FERREIRA

REDUÇÃO DA DIMENSIONALIDADE PARA ESTIMATIVA DE TEORES DE
NUTRIENTES EM FOLHAS E GRÃOS DE SOJA COM ESPECTROSCOPIA NO
INFRAVERMELHO

Dissertação apresentada ao Programa de
Pós-Graduação em Computação Aplicada,
curso de Mestrado em Computação Aplicada
da Universidade Estadual de Ponta Grossa,
como requisito para obtenção do título de
Mestre.

Orientação: Prof^ª. Dra. Alaine Margarete
Guimarães

Coorientação: Prof. Dr. Eduardo Fávero
Caires

PONTA GROSSA
2017

Ficha Catalográfica
Elaborada pelo Setor de Tratamento da Informação BICEN/UEPG

F383 Ferreira, Pablo Henrique
Redução da dimensionalidade para
estimativa de teores de nutrientes em
folhas e grãos de soja com espectroscopia
no infravermelho/ Pablo Henrique Ferreira.
Ponta Grossa, 2017.
93f.

Dissertação (Mestrado em Computação
Aplicada - Área de Concentração:
Computação para Tecnologias em
Agricultura), Universidade Estadual de
Ponta Grossa.

Orientadora: Prof^a Dr^a Elaine Margarete
Guimarães.

Coorientador: Prof. Dr. Eduardo Fávero
Caires.

1.Refletância espectral. 2.Mineração de
dados. 3.Seleção de variáveis. 4.Modelos
de regressão. I.Guimarães, Elaine
Margarete. II. Caires, Eduardo Fávero.
III. Universidade Estadual de Ponta
Grossa. Mestrado em Computação Aplicada.
IV. T.

CDD: 004.3

TERMO DE APROVAÇÃO

Pablo Henrique Ferreira

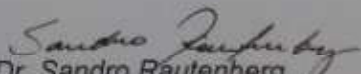
"REDUÇÃO DA DIMENSIONALIDADE PARA ESTIMATIVA DE TEORES DE NUTRIENTES EM FOLHAS E GRÃOS DE SOJA COM ESPECTROSCOPIA NO INFRAVERMELHO".

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Orientadora:


Dr.ª Alaine Margarete Guimarães
UEPG


Dr. José Carlos Ferreira da Rocha
UEL


Dr. Sandro Rautenberg
UNICENTRO/Guarapuava - PR

Ponta Grossa, 27 de abril de 2017.

*Aos meus pais, minhas irmãs e à
Marines Steinmetz por serem meu porto seguro
perante as dificuldades durante o percurso.*

AGRADECIMENTOS

A Deus por ter me dado sabedoria, saúde e paz durante toda minha vida.

Agradeço a minha família pelo apoio e pelos conselhos durante o desenvolvimento deste trabalho, sempre presente nos momentos críticos e me apoiando a sempre continuar nesse caminho.

A minha namorada Marines Steinmetz pelos momentos, força, apoio e por estar presente em todos os momentos que precisei, sempre me dando forças para continuar nessa caminhada.

Aos meus futuros padrinhos de casamento Cristina e Vladimir pelas conversas, conselhos e apoio sempre dispostos a oferecer ajuda com maior carinho e dedicação.

A minha orientadora Profa. Dra. Alaine Margarete Guimarães pela dedicação, tempo e oportunidade a elaborar este trabalho, pelo apoio e incentivo nos momentos difíceis desta caminhada.

Ao meu coorientador Prof. Dr. Eduardo Fávero Caires e ao Laboratório de Solos da UEPG pela disponibilização das amostras de folhas e grãos de soja, cuja importância foi essencial neste trabalho.

Aos amigos de infância Robson Felliipe, Ubaldo Seglin Neto e Matheus Oliveira pela amizade e companheirismo que levarei por toda minha vida. Aos meus primos Edson Cunha, Luis Guilherme e Lucas Cunha pela amizade e descontração nos momentos difíceis.

Aos amigos do INFOAGRO, Victor, Rayson, Jackson, Perin, David e Alessandro pelo companheirismo e apoio a encarar e resolver os desafios desta caminhada.

Aos amigos que fiz durante o Mestrado e todos que participaram direta e indiretamente deste trabalho, pelo incentivo e apoio.

A todos os professores do curso de Mestrado em Computação Aplicada que dedicaram-se em transmitir seus conhecimentos que me ajudaram imensamente na elaboração deste trabalho.

A Universidade Estadual de Ponta Grossa e ao programa de Pós-Graduação em

Computação Aplicada que forneceram todo suporte e infraestrutura para desenvolvimento desta dissertação.

Ao Laboratório Multiusuário disponibilizando todos os recursos necessários para esta pesquisa, como também a Simone Sabino e a todos os integrantes do C-LABMU pelo ensino e colaboração durante a etapa de análises das amostras.

Por fim, agradeço a CAPES pelo apoio financeiro essencial durante os dois anos de trabalho.

Muito Obrigado!

RESUMO

A alta dimensionalidade em bases de dados é um problema que pode estar presente em diversos segmentos, inclusive nas análises do estado de nutrientes em plantas. Atualmente essas análises são baseadas em metodologias que demandam tempo e reagentes. A espectroscopia do infravermelho próximo (NIR – *NearInfrared*) e médio (MIR – *MiddleInfrared*) têm se mostrado uma alternativa mais rápida e limpa em relação a quantificação simultânea de compostos. Os dados obtidos por esses equipamentos apresentam alta dimensão. A leitura ocorre em comprimentos de onda gerando centenas atributos para o NIR e milhares para o MIR. Uma das dificuldades está em identificar quais atributos são mais relevantes para análise dos nutrientes. Este trabalho teve como objetivo verificar o ganho de correlação obtido com o emprego de redução de dimensionalidade em dados obtidos por espectroscopia NIR e MIR, para estimativa de teores de 11 nutrientes em grãos e folhas de soja, sendo eles: Nitrogênio (N), Fósforo (P), Potássio (K), Cálcio (Ca), Magnésio (Mg), Enxofre (S), Cobre (Cu), Manganês (Mn), Ferro (Fe), Zinco (Zn) e Boro (B). Para isto, 231 amostras de folhas de soja e 285 de grãos de soja foram utilizadas para geração de modelos de regressão, sendo os espectros obtidos através dos espectrofotômetros NIR e MIR. Os modelos de regressão foram gerados pelos algoritmos de aprendizado de máquina SMOREg que implementa a máquina de vetor de suporte para regressão, o algoritmo baseado em árvores de decisão com funções de regressão M5Rules e o algoritmo LinearRegression. Os resultados foram avaliados através do coeficiente de correlação (r) e o erro quadrático (RRSE). A estimativa de nutrientes para folhas foi satisfatória tanto para espectroscopia NIR e MIR, onde correlações acima de 0,80 foram obtidas para os nutrientes P, K, Mg, S, Mn, Cu, Fe e Zn. Não houve correlações para B e Ca em folhas de soja. A estimativa de teores de nutrientes foi também satisfatória para grãos de soja, mas apenas em dados de espectroscopia NIR, onde correlações acima de 0,7 foram obtidas para N, P, K, Ca e S. O uso da redução de dimensionalidade proporcionou os altos valores para correlação de P, K e S em folhas de soja, fazendo uso do algoritmo LinearRegression. Para os grãos de soja, a redução de dimensionalidade foi imprescindível na obtenção de correlações satisfatórias, exceto para N, sempre utilizando o algoritmo LinearRegression. Quando a redução da dimensionalidade não foi usada, os resultados satisfatórios foram obtidos pelo algoritmo SMOREg a partir de dados foliares para os nutrientes N, Mg, Cu, Mn, Fe e Zn. A utilização da redução de dimensionalidade junto ao algoritmo LinearRegression auxiliou na obtenção de melhores correlações para três nutrientes em folhas e para os índices satisfatórios de grãos. Os resultados observados demonstram uma maior eficiência no uso do NIR para análises foliares do que para análises de grãos. As técnicas computacionais SMOREg e LinearRegression obtiveram os melhores resultados, sendo a SMOREg indicada para grandes quantidades de atributos e LinearRegression para quantidades menores de atributos.

Palavras-chave: Refletância Espectral, Mineração de Dados, Seleção de Variáveis, Modelos de Regressão

ABSTRACT

The high dimensionality in databases is a problem that can occur in several fields, including the plants nutrients state analysis. These analyses are currently based on methodologies that spend time and reagents. (NIR-*NearInfrared*) and (MIR-*MiddleInfrared*) spectroscopy have been shown to be a faster and clean alternative to simultaneous quantification of compounds. Since reading occurs at wavelengths generating hundreds attributes for the NIR and thousands to the MIR the data obtained by such equipment have a high dimensionality. One of the difficulties is to identify which attributes are more relevant for the nutrient analysis. This work aimed to verify the correlation gain obtained with the use of dimensionality reduction techniques with data obtained by NIR and MIR spectroscopy. The goal is to estimated levels of 11 nutrients in grains and leaves of soybean: Nitrogen (N), Phosphorus (P), Potassium (K), Calcium (Ca), Magnesium (Mg), Sulfur (S), Copper (Cu), Manganese (Mn), Iron (Fe), Zinc (Zn) and Boron (B). For that, 231 soybean leaves and 285 soybeans samples were analysed by spectroscopy in the mid-infrared and near-infrared region. The regression models were generated by machine learning algorithms: SMOREg which implements the support vector machine for regression; M5Rules that is based on decision trees with regression functions; and LinearRegression algorithm for linear regression. The results were evaluated by correlation coefficient (r) and the quadratic error (RRSE). Estimating leaf nutrients was satisfactory for both NIR and MIR spectroscopy, where correlations of 0.80 above were obtained for P, K, Mg, S, Mn, Cu, Fe and Zn. There were no correlations for B and Ca in soybean leaves. Estimating nutrient was also satisfactory for soybeans, but only in NIR spectroscopy data, where correlations above 0.7 were obtained for N, P, K, Ca, and S. Using dimensionality reduction techniques provided the high values for correlation of P, K, and S in soybean leaves, making use of the LinearRegression algorithm. For soybeans, the dimensionality reduction was essential in obtaining satisfactory correlations, except for N, always using the LinearRegression algorithm. When reducing the dimensionality was not used, satisfactory results were obtained by the SMOREg algorithm from foliar data to N, Mg, Cu, Mn, Fe, and Zn. Reducing dimensionality associated to the use of LinearRegression algorithm resulted in better correlations for three nutrients in leaves and satisfactory rates of grain. The observed results demonstrate a greater efficiency in the use of the NIR for foliar analysis than for grain analysis. SMOREg computational techniques and LinearRegression algorithm presented the best results, being the SMOREg indicated for large quantities of attributes and LinearRegression for smaller quantities.

Keywords: Spectral Reflectance, Data Mining, Selection of Variables, Regression Models

LISTA DE ILUSTRAÇÕES

Figura 1. Espectrofotômetro NIR, marca FOSS, modelo DA 1650, onde foram lidas as amostras de folhas e grãos.	37
Figura 2. Amostra com 40 ml de folha de soja no acessório SmallCup do espectrofotômetro NIR.	37
Figura 3. Espectrofotômetro MIR, marca Shimadzu, modelo IRPrestige-21, utilizado para leituras de amostras de folhas.	38
Figura 4. Descrição de pastilha feita da amostra de folha de soja com KBr (Figura 4(a)) e Pastilhador da marca Shimadzu (Figura 4(b)).	38
Quadro 1. Descrição dos 28 atributos utilizados na execução dos algoritmos gerando correlações acima de 0,70 no intervalo de 1.128 a 1.568 nm. ...	43
Quadro 2. Descrição dos 6 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,69 no intervalo de 1.404 a 1.500 nm.	43
Quadro 3. Descrição dos 36 atributos utilizados na execução dos algoritmos LinearRegression e M5Rules gerando correlações acima de 0,70 e também 16 atributos pelo qual os dois algoritmos obtiveram correlações de 0,73 e 0,84 respectivamente no intervalo de 1.104 a 1.642 nm.	44
Quadro 4. Descrição dos 31 atributos utilizados na execução dos algoritmos gerando correlações acima de 0,70 no intervalo de 1.102 a 1.600 nm.	46
Quadro 5. Descrição dos 9 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,68 no intervalo de 1.100 a 1.182 nm.	47
Quadro 6. Descrição dos 14 atributos utilizados na execução do algoritmos LinearRegression gerando correlação de 0,74 no intervalo de 1.178 a 1.638 nm.	48
Quadro 7. Descrição dos 15 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,75 no intervalo de 1.100 a 1.566 nm. ...	48
Quadro 8. Descrição dos 23 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,75 no intervalo de 407 a 3.584 cm-1. ...	54

Quadro 9. Descrição dos 15 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,70 no intervalo de 1.102 a 1.338 nm.	57
Quadro 10. Descrição dos 27 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,85 no intervalo de 1.100 a 1.486 nm.	58
Quadro 11. Descrição dos 15 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,75 no intervalo de 1.180 a 1.480 nm.	58
Quadro 12. Descrição dos 10 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,76 no intervalo de 1.182 a 1.478 nm.	59
Quadro 13. Descrição dos 24 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,84 no intervalo de 1.100 a 1.498 nm.	60
Quadro 14. Descrição dos 17 atributos utilizados na execução dos algoritmos LinearRegression e M5Rules gerando correlações de 0,87 e 0,84, respectivamente, no intervalo de comprimentos de onda de 1.258 a 1.534 nm.	65
Quadro 15. Descrição dos 19 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,90 no intervalo de comprimentos de onda de 1.102 a 1.536 nm.	66
Quadro 16. Descrição dos 28 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,86 no intervalo de comprimentos de onda de 1.240 a 1.542 nm.	68
Quadro 17. Descrição dos 27 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,70 no intervalo de comprimentos de onda de 1.100 a 1.158 nm.	69
Quadro 18. Descrição dos 53 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,81 no intervalo de números de onda de 399 a 3628cm-1.	76
Quadro 19. Descrição dos 12 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,70 no intervalo de comprimentos de onda de 1.102 a 1.322 nm.	79

Quadro 20. Descrição dos 16 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,71 no intervalo de comprimentos de onda de 1.104 a 1.478 nm. 80

Quadro 21. Descrição dos 24 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,77 no intervalo de comprimentos de onda de 1.118 a 1.486 nm. 81

Quadro 22. Descrição dos 10 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,73 no intervalo de comprimentos de onda de 1.206 a 1.454 nm. 81

Quadro 23. Descrição dos 21 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,78 no intervalo de comprimentos de onda de 1.102 a 1.476 nm. 82

LISTA DE TABELAS

Tabela 1. Regiões Espectrais do Infravermelho	26
Tabela 2. Descrição de teores de nutrientes obtidas a partir das análises de laboratório para folhas e grãos de soja, como também os valores mínimos, máximos e média. . . .	39
Tabela 3. Resultados da análise foliar utilizando redução da dimensionalidade em 10 % dos atributos.	41
Tabela 4. Resultados da análise foliar utilizando redução da dimensionalidade em 20 % dos atributos.	42
Tabela 5. Resultados da análise foliar sem utilizar redução da dimensionalidade, totalizando 275 atributos.	42
Tabela 6. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de N com diferentes quantidades de atributos.	43
Tabela 7. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.	43
Tabela 8. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.	44
Tabela 9. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.	45
Tabela 10. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.	45
Tabela 11. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.	46
Tabela 12. Resultados de correlações obtidos pelos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.	46
Tabela 13. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.	47
Tabela 14. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Fe com diferentes quantidades de atributos.	48

Tabela 15. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.	48
Tabela 16. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.	49
Tabela 17. Resultados da análise foliar utilizando redução da dimensionalidade em 10 % dos atributos.	49
Tabela 18. Resultados da análise foliar utilizando redução da dimensionalidade em 20 % dos atributos.	50
Tabela 19. Resultados da análise foliar sem utilizar redução da dimensionalidade, totalizando 1868 atributos.	50
Tabela 20. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de N com diferentes quantidades de atributos.	51
Tabela 21. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.	51
Tabela 22. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.	51
Tabela 23. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.	52
Tabela 24. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.	52
Tabela 25. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.	53
Tabela 26. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.	53
Tabela 27. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.	53
Tabela 28. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Fe com diferentes quantidades de atributos.	54
Tabela 29. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.	54

Tabela 30. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.	55
Tabela 31. Resultados da análise de grãos utilizando redução da dimensionalidade em 10 % dos atributos.	55
Tabela 32. Resultados da análise de grãos utilizando redução da dimensionalidade em 20 % dos atributos.	56
Tabela 33. Resultados da análise de grãos sem utilizar redução da dimensionalidade, totalizando 275 atributos.	56
Tabela 34. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de N com diferentes quantidades de atributos.	57
Tabela 35. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.	57
Tabela 36. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.	58
Tabela 37. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.	59
Tabela 38. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.	59
Tabela 39. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.	60
Tabela 40. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.	60
Tabela 41. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.	61
Tabela 42. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Fe com diferentes quantidades de atributos.	61
Tabela 43. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.	61
Tabela 44. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.	62

Tabela 45. Resultados da análise foliar em dados transformados com redução da dimensionalidade em 10 % dos atributos.	63
Tabela 46. Resultados da análise foliar em dados transformados com redução da dimensionalidade em 20 % dos atributos.	63
Tabela 47. Resultados da análise foliar em dados transformados sem utilizar redução da dimensionalidade, totalizando 825 atributos.	64
Tabela 48. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de N com diferentes quantidades de atributos.	64
Tabela 49. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.	65
Tabela 50. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.	66
Tabela 51. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.	66
Tabela 52. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.	67
Tabela 53. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.	67
Tabela 54. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.	68
Tabela 55. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.	69
Tabela 56. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Fe com diferentes quantidades de atributos.	69
Tabela 57. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.	70
Tabela 58. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.	70
Tabela 59. Resultados da análise foliar em dados transformados com redução da dimensionalidade em 10 % dos atributos.	71

Tabela 60. Resultados da análise foliar em dados transformados com redução da dimensionalidade em 20 % dos atributos.	71
Tabela 61. Resultados da análise foliar em dados transformados sem utilizar redução da dimensionalidade, totalizando 5604 atributos.	72
Tabela 62. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de N com diferentes quantidades de atributos.	72
Tabela 63. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.	73
Tabela 64. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.	73
Tabela 65. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.	73
Tabela 66. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.	74
Tabela 67. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.	74
Tabela 68. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.	75
Tabela 69. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.	75
Tabela 70. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativas do teor de Fe com diferentes quantidades de atributos.	75
Tabela 71. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.	76
Tabela 72. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.	77
Tabela 73. Resultados da análise de grãos em dados transformados com redução da dimensionalidade em 10 % dos atributos.	78
Tabela 74. Resultados da análise de grãos em dados transformados com redução da dimensionalidade em 20 % dos atributos.	78

Tabela 75. Resultados da análise de grãos sem utilizar redução da dimensionalidade, totalizando 825 atributos.	79
Tabela 76. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativas do teor de N com diferentes quantidades de atributos.	79
Tabela 77. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.	80
Tabela 78. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.	80
Tabela 79. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.	81
Tabela 80. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.	82
Tabela 81. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.	82
Tabela 82. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.	83
Tabela 83. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.	83
Tabela 84. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Fe com diferentes quantidades de atributos	84
Tabela 85. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.	84
Tabela 86. Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.	84

LISTA DE SIGLAS

ABTS	2,2-azinobis-3-ethylbenz-thiazoline-6-sulfonic acid
DPPH	2,2-difenil-1-picril-hidrazila (do inglês, <i>2,2-diphenyl-1-picrylhydrazyl</i>)
FIR	Infravermelho Distante (do inglês, <i>Far Infrared</i>)
FRAP	Capacidade de redução férrica (do inglês, <i>Ferric reducing/antioxidant power</i>)
IA	Inteligência Artificial
IUPAC	<i>International Union of Pure and Chemistry</i>
LS-SVM	Máquina de Vetor de Suporte por Mínimos Quadrados
MD	Mineração de Dados
MIR	Infravermelho Médio (do inglês, <i>Middle Infrared</i>)
MLR	Regressão Linear Múltipla (do inglês, <i>Multiple Linear Regression</i>)
MO	Matéria Orgânica
MPLS	Mínimos Quadrados Parciais Modificada (do inglês, <i>Modified Partial Least-Squares</i>)
NDVI	Índice de Vegetação por Diferença Normalizada
NIR	Infravermelho próximo (do inglês, <i>Near Infrared</i>)
NIRS	Espectroscopia no Infravermelho Próximo (do inglês, <i>Near Infrared Spectroscopy</i>)
PCA	<i>Principal Component Analysis</i>
PCR	Regressão em Componentes Principais (do inglês, <i>Principal Components Regression</i>)
PLS	Regressão por Quadrados Mínimos Parciais (do inglês, <i>Partial Least Squares</i>)
R ²	Coeficiente de Determinação
r	Coeficiente de Correlação
RRSE	Relative Root Squared Error
SVD	<i>Singular Value Decomposition</i>
SVMR	Regressão de Máquina de Vetor de Suporte (do inglês, <i>Support Vector Machine Regression</i>)
Vis-NIR	Visível - Infravermelho Próximo (do inglês, <i>Visible-Near Infrared</i>)

SUMÁRIO

1	Introdução	22
2	Objetivos	25
2.1	Objetivo Geral	25
2.2	Objetivos Específicos	25
3	Revisão de Literatura	26
3.1	Espectroscopia no Infravermelho Próximo (NIRS) e Médio (MIRS)	26
3.2	Mineração de Dados	30
3.2.1	Seleção de Atributos para Redução de Dimensionalidade	31
3.3	Algoritmos para Tarefa de Regressão	34
4	Material e Métodos	36
4.1	Área de Coleta de Amostras de Folha e Grãos de Soja	36
4.2	Preparo e Análise - Equipamento NIRS e MIRS	36
4.3	Elaboração da Base de Dados	39
4.4	Seleção de Atributos	40
4.5	Mineração de Dados – Determinação de Modelos de Regressão	40
5	Resultados	41
5.1	Análise Foliar	41
5.1.1	Espectroscopia NIRS	41
5.1.1.1	N - Nitrogênio	42
5.1.1.2	P - Fósforo	43
5.1.1.3	K - Potássio	44
5.1.1.4	Ca - Cálcio	44
5.1.1.5	Mg - Magnésio	45

5.1.1.6	S - Enxofre	45
5.1.1.7	Cu - Cobre	46
5.1.1.8	Mn - Manganês	47
5.1.1.9	Fe - Ferro	47
5.1.1.10	Zn - Zinco	48
5.1.1.11	B - Boro	48
5.1.2	Espectroscopia MIRS	49
5.1.2.1	N - Nitrogênio	50
5.1.2.2	P - Fósforo	51
5.1.2.3	K - Potássio	51
5.1.2.4	Ca - Cálcio	52
5.1.2.5	Mg - Magnésio	52
5.1.2.6	S - Enxofre	52
5.1.2.7	Cu - Cobre	53
5.1.2.8	Mn - Manganês	53
5.1.2.9	Fe - Ferro	54
5.1.2.10	Zn - Zinco	54
5.1.2.11	B - Boro	55
5.2	Análise de Grãos	55
5.2.1	N - Nitrogênio	56
5.2.2	P - Fósforo	57
5.2.3	K - Potássio	58
5.2.4	Ca - Cálcio	58
5.2.5	Mg - Magnésio	59
5.2.6	S - Enxofre	59
5.2.7	Cu - Cobre	60

5.2.8	Mn - Manganês	60
5.2.9	Fe - Ferro	61
5.2.10	Zn - Zinco	61
5.2.11	B - Boro	62
5.3	Análise Foliar - Dados Transformados	62
5.3.1	Espectroscopia NIR	62
5.3.1.1	N - Nitrogênio	64
5.3.1.2	P - Fósforo	65
5.3.1.3	K - Potássio	65
5.3.1.4	Ca - Cálcio	66
5.3.1.5	Mg - Magnésio	67
5.3.1.6	S - Enxofre	67
5.3.1.7	Cu - Cobre	68
5.3.1.8	Mn - Manganês	68
5.3.1.9	Fe - Ferro	69
5.3.1.10	Zn - Zinco	69
5.3.1.11	B - Boro	70
5.3.2	Espectroscopia MIR	70
5.3.2.1	N - Nitrogênio	72
5.3.2.2	P - Fósforo	72
5.3.2.3	K - Potássio	73
5.3.2.4	Ca - Cálcio	73
5.3.2.5	Mg - Magnésio	74
5.3.2.6	S - Enxofre	74
5.3.2.7	Cu - Cobre	74
5.3.2.8	Mn - Manganês	75

5.3.2.9	Fe - Ferro	75
5.3.2.10	Zn - Zinco	76
5.3.2.11	B - Boro	76
5.4	Análise de Grãos - Dados Transformados	77
5.4.1	N - Nitrogênio	79
5.4.2	P - Fósforo	80
5.4.3	K - Potássio	80
5.4.4	Ca - Cálcio	81
5.4.5	Mg - Magnésio	81
5.4.6	S - Enxofre	82
5.4.7	Cu - Cobre	82
5.4.8	Mn - Manganês	83
5.4.9	Fe - Ferro	83
5.4.10	Zn - Zinco	84
5.4.11	B - Boro	84
6	Discussão	85
6.1	Análise Foliar	85
6.2	Análise de Grãos	86
7	Conclusão	88
	REFERÊNCIAS	90

1 INTRODUÇÃO

Nos últimos 40 anos, o Brasil tem passado por uma grande evolução no que diz respeito a produção agrícola, tanto em relação a aspectos técnicos como comerciais graças ao uso correto e eficiente de insumos (sementes, adubos e máquinas). A melhoria da produtividade das lavouras de grãos tem contribuído significativamente para essa evolução. Porém, para gerar retornos satisfatórios em relação ao investimento aplicado, tais insumos devem ter sua aplicação otimizados de maneira a maximizar os retornos. A utilização correta necessita de uma diagnose precisa em relação aos problemas do solo e da nutrição das plantas, a qual é normalmente obtida por meio de análises laboratoriais.

As análises em laboratórios têm aumentado na mesma proporção que a produção agrícola. A análise de solo como ferramenta de diagnose tem sido avaliada como a base de um programa de adubação. A análise foliar é fundamental na fase de adubação, permitindo a identificação de deficiências nutricionais. Complementarmente, a análise de grãos pode auxiliar durante o programa de adubação.

Os métodos utilizados de análises de solo em laboratórios demandam um conjunto de procedimentos implicando em um tempo de análise que pode levar dias. Também, esses métodos requerem o uso de reagentes químicos, cujo uso causa riscos a saúde humana e ao meio ambiente. Novos procedimentos analíticos vêm sendo estudados e aplicados, entre eles o emprego da técnica de infravermelho próximo (NIR - *Near Infrared*) e infravermelho médio (MIR - *Middle Infrared*). A espectroscopia do infravermelho efetua análises de moléculas orgânicas através do princípio de radiação eletromagnética, no qual incide-se uma radiação sobre uma amostra moída ou líquida, obtendo uma quantidade de energia capaz de vibrar as ligações químicas que unem os átomos. O resultado desta operação pode ser transformado matematicamente na concentração do nutriente que se deseja determinar (NETO, 2005).

Quando a amostra é irradiada, a luz é absorvida e assim dá-se origem a um espectro. Tais espectros são obtidos através de espectrofotômetros que apresentam diferentes características com base em seus componentes e também em relação a faixa de comprimento de onda de operação (visível, infravermelho) e a forma em que as medidas espectrofotométricas (transmitância ou refletância) são obtidas (SALIBA, 1999). Os espectros de uma molécula são constituídos por bandas e a energia associada a ela é a soma das energias eletrônicas, vibracionais e rotacionais. Na região do infravermelho são observadas as energias vibracionais (SCHRADER, 2008), sendo que a região do espectro infravermelho próximo - NIR corresponde aos comprimentos de onda de 780 até 2.500 nm e região do

infravermelho médio - MIR de 2.500 até 25.000 nm (SKOOG; HOLLER; NIEMAN, 2002).

As técnicas de infravermelho NIRS e MIRS podem vir a substituir boa parte das metodologias convencionais de análises de laboratório. Possui grandes vantagens, pois pode analisar várias características por amostra em um tempo médio de um minuto, requer pouca mão-de-obra, além de não poluir o meio ambiente por não utilizar produtos químicos ou reagentes (AMORIM, 1996). Entretanto, as desvantagens em trabalhar com esse sistema consistem em seu alto custo de aquisição e a necessidade em algumas situações de grande quantidade de material a ser analisado para que seja determinada uma calibração segura referente a variável de interesse. Neste sentido, mesmo sendo uma técnica de análise avançada, sua efetividade ainda depende dos métodos analíticos convencionais, pois a calibração é feita em função dos resultados destes.

Considerando a quantidade de dados necessários para calibração e consequente geração de modelos de resposta espectral para diferentes elementos do solo e da folha, métodos computacionais de análise desses dados se tornam importantes na obtenção de resultados eficientes. Essa eficiência não é de fácil obtenção, visto que os dados apresentam alta dimensionalidade além da similaridade entre variáveis, aqui chamadas de atributos. Essa alta dimensionalidade ocorre em função de cada comprimento de onda lido corresponder a um atributo. Logo, considerando-se o espectro do infravermelho tem-se milhares de atributos. A similaridade entre os atributos deve-se ao fato de que todos correspondem a um mesmo tipo de dado, e a resposta espectral em diferentes comprimentos de onda. Além disso, no espectro, vários comprimentos de onda têm respostas muito semelhantes para um mesmo elemento sendo analisado.

As técnicas avançadas de mineração de dados (MD) tem se mostrado úteis na análise de dados de alta dimensão e de alta similaridade. Sendo o objetivo principal da mineração de dados a descoberta de conhecimento potencialmente útil em grandes e complexas bases de dados, seus métodos podem, dessa forma, contribuir na identificação de um modelo de calibração que permita a estimativa de teor de nutrientes do solo e das folhas por meio de refletância espectral.

Em tais situações, deve-se aplicar algoritmos de mapeamento e seleção que produzem uma representação desses dados em dimensões menores. Na literatura existem diversos algoritmos de MD como a PCA - *Principal Component Analysis* que tratam destes problemas (TAN; STEINBACH; KUMAR, 2006). Técnicas de redução da dimensionalidade usadas em MD visam reproduzir um conjunto de dados de dimensão determinada em outro espaço de dimensão menor e que sejam mantidas as mesmas características do

conjunto.

Assim, a determinação de modelos de calibração de dados de refletância espectral na região do NIR e MIR, visando a estimativa de nutrientes em folhas e grãos de soja, utilizando técnicas adequadas de seleção de atributos e mineração de dados, mostra-se relevante e corresponde ao tema de estudo desse trabalho.

2 OBJETIVOS

2.1 Objetivo Geral

Considerando a hipótese de que a redução de dimensionalidade em dados de espectroscopia leva a geração de modelos mais precisos, o objetivo desse trabalho foi verificar o ganho de correlação obtido com o emprego de redução de dimensionalidade em dados obtidos por espectroscopia NIR e MIR, para estimativa de teores de nutrientes em folhas e grãos de soja.

2.2 Objetivos Específicos

- Estimar teores de nutrientes em folhas e em grãos de soja, com base nos dados obtidos por espectroscopia NIR e MIR;
- Avaliar diferentes algoritmos de regressão em relação a capacidade de estimativa de nutrientes em folhas e grãos de soja em bases de dados com e sem redução de dimensionalidade;
- Avaliar qual técnica de espectroscopia (NIR e MIR) estima com mais eficiência os nutrientes em folhas e grãos de soja.

3 REVISÃO DE LITERATURA

3.1 Espectroscopia no Infravermelho Próximo (NIRS) e Médio (MIRS)

O termo espectroscopia consiste no estudo da radiação eletromagnética emitida ou absorvida por um corpo. A região do espectro que corresponde ao infravermelho está localizada depois da região do visível e esta dividida em infravermelho próximo - NIR (*Near-Infrared*), infravermelho médio - MIR (*Middle-Infrared*) e infravermelho distante - FIR (*Far-Infrared*) (Tabela 1) (SKOOG; HOLLER; NIEMAN, 2002).

Tabela 1: Regiões Espectrais do Infravermelho

Região	Número de onda (ν) cm^{-1}	Região em comprimento de onda (λ), nm
Próximo (NIR)	12.800 a 4.000	780 a 2.500
Médio (MIR)	4.000 a 400	2.500 a 25.000
Distante (FIR)	400 a 10	25.000 a 1.000.000

Fonte: (SKOOG; HOLLER; NIEMAN, 2002)

Não há consenso na literatura sobre a delimitação de uma região para outra. Pasquini (2003) define a região do NIR entre 750 a 2.500 nm, no entanto, a IUPAC (*International Union of Pure and Chemistry*) define a região do NIR sendo entre 780 a 2.500 nm e MIR entre 2.500 a 25.000 (DAVIES, 1998).

A espectroscopia no infravermelho próximo NIRS (do inglês, *Near InfraRed Spectroscopy*) e médio MIRS (do inglês, *Middle InfraRed Spectroscopy*) consistem de técnicas não-destrutivas utilizadas para avaliação de materiais, sendo que esta técnica requer pouco tratamento na fase de preparação das amostras, e em alguns casos nenhum. É baseada na espectroscopia vibracional, que mede a interação da radiação eletromagnética com a matéria (PASQUINI, 2003).

A espectroscopia no infravermelho é utilizada desde a década de 60, cujo objetivo era determinar umidade em sementes de várias culturas, tais como trigo, milho, aveia, cevada, sorgo e arroz (HART; NORRIS; GOLUMBIC, 1962). Tem sido utilizada cada vez mais como método analítico, pois tem atendido a demanda de análises nos laboratórios com eficiência. Para as regiões do infravermelho, têm sido realizados trabalhos com amostras de diferentes estados e formas como gases, sólidos e líquidos, como também em amostras semissólidas, pastas, géis e outras (SCHRADER, 2008). Aplicações das técnicas NIRS e MIRS têm fornecido resultados adequados para determinação de moléculas orgânicas, e na área de nutrição de plantas o sistema NIR tem sido usado na determinação de nutri-

entes em folhas de cana-de-açúcar (SANTOS; PEREIRA; KORNDÖRFER, 2010) e também de nitrogênio em folhas de trigo (LIMA *et al.*, 2008).

Wu *et al.* (2012) realizaram um estudo para avaliar a viabilidade em utilizar a técnica de espectroscopia (NIRS) para a determinação do índice de atividade em três tipos de antioxidantes em extratos de folhas de bambu, especificamente DPPH (2,2-diphenyl-1-picrylhydrazyl), FRAP (*ferric reducing/antioxidant power*) e o ABTS (2,2-azinobis-3-ethylbenz-thiazoline-6-sulfonic acid). Os modelos de calibração estabelecidos pela regressão linear múltipla, apresentaram coeficientes de correlação de 0,86, 0,91 e 0,96, respectivamente.

Anamari *et al.* (2012) buscaram avaliar a resposta espectral do dossel de plantas à variação de doses de nitrogênio (N) e sua relação com os teores foliares de N, o índice de clorofila e a produtividade na cultura do algodoeiro. Os dados espectrais foram obtidos a partir do sensor ótico ativo *Crop Circle*® ACS-210, o qual emite a luz no comprimento de onda da região do visível a 580 nm e do infravermelho próximo a 880nm. A partir da refletância, esse equipamento calcula o índice de vegetação por diferença normalizada (NDVI). O índice de clorofila (ou “índice verde”) presente na folha foi medido a partir do clorofilômetro ClorofiLOG CFL 1030 nos comprimentos de onda de 650 e 940 nm. As maiores correlações obtidas dos índices de clorofila e do NDVI com a produtividade foram observadas aos 56 dias após a emergência e antes da senescência. Como resultado foi indicado que o sensor tornou-se mais sensível à variação do teor foliar de N conforme a planta se desenvolve.

Santos, Santos e Korndörfer (2012) utilizaram da técnica NIRS para estimar valores de nitrogênio foliar de milho e soja. Para construção de um banco de dados e curvas de calibração, foram selecionados 100 espectros dos 315 obtidos da leitura em triplicata de 105 amostras de folhas de milho e 73 espectros dos 363 resultantes da leitura em triplicata de 121 amostras de folhas de soja. Na etapa de validação das curvas os autores obtiveram coeficientes de correlação de 0,85 para folhas de milho e 0,88 para soja, sendo que os resultados estimados não diferiram estatisticamente do método convencional. Na etapa de predição, utilizando-se de 92 amostras de folhas de milho e 86 amostras de folhas de soja que não pertenciam a base de dados analisada, os valores encontrados pelo NIRS não diferiram dos resultados do método convencional por laboratório, e obtiveram coeficientes de correlação de 0,79 e 0,85 para milho e soja respectivamente, indicando assim que a técnica NIRS expressa o teor de nitrogênio foliar com confiabilidade.

O trabalho realizado por Haughey *et al.* (2013) procurou distinguir em amostras

de soja, quais estavam adulteradas com melanina. A partir de tais amostras, foram gerados espectros NIR e obteve-se coeficientes de determinação entre 0,89 e 0,99 dependendo do algoritmo matemático utilizado e do pré-processamento das amostras. Os autores encontraram valores correspondentes para o erro quadrático médio de calibração e predição de 0,081-0,276% e 0,134-0,368%, respectivamente. Além disso a utilização de uma abordagem qualitativa com os dados espectrais e a aplicação da técnica PCA (*Principal Component Analysis*) possibilitou distinguir entre as amostras quais eram adulteradas e não adulteradas.

Santos, Pereira e Korndörfer (2010) avaliaram o equipamento NIR na determinação do teor de matéria orgânica (M.O.), argila de solos e teores foliares de silício (Si) e nitrogênio (N). A base de dados foi composta por 122 amostras de diferentes tipos de solos e 47 amostras de folhas de cana-de-açúcar de sete variedades obtidas pelo laboratório da usina de cana-de-açúcar “Jalles Machado” situada no município de Goianésia/GO. Os resultados obtidos pelo NIRS, e os obtidos pelos métodos analíticos de referência mostraram coeficientes de correlação de 0,87, 0,83, 0,77 e 0,70, respectivamente para os teores foliares de N e Si e, argila e MO no solo, indicando uma maior eficiência do NIRS para análises foliares. Os autores concluíram que para os quatro tipos de análises, não houve diferença entre os resultados das análises realizadas pelo método NIRS daqueles realizados pelos métodos convencionais.

Souza, Madari e Guimarães (2012) realizaram um trabalho com objetivo de determinar matéria orgânica em amostras de solo aplicando técnicas de Inteligência Artificial (IA) na análise de espectros do infravermelho médio (MIR), obtendo valores para a técnica LS-SVM (*Least Square Support Vector Machine*) em que o Erro Quadrático Médio de Previsão foi $3,26 \text{ g/Kg}^{-1}$ em comparação ao método de Walkley-Black de $2,80 \text{ g/Kg}^{-1}$. Os autores citam que durante o ano de 2010 foram realizadas 5000 determinações de matéria orgânica em solos no Laboratório de Análise Agroambiental (LAA) do Centro Nacional de Pesquisa de Arroz e Feijão da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), em Goiás. Isso corresponde ao consumo de 2,5 Kg de dicromato de potássio ($\text{K}_2\text{Cr}_2\text{O}_7$, grau analítico para análise - P.A.), 100 L de ácido sulfúrico (H_2SO_4) concentrado e geração de aproximadamente 1400 L de resíduo sulfocrômico. Tais desvantagens descritas justificam esforços no desenvolvimento cada vez maior de metodologias limpas, como a análise por espectroscopia no infravermelho. Os resultados encontrados no trabalho demonstram que a aplicação da técnica MIRS na determinação de MO apresenta baixo custo das análises e otimização da operacionalidade.

Masoni, Ercoli e Mariotti (1996) utilizaram a técnica de espectroscopia do in-

fravermelho próximo e do visível para detectar deficiências de minerais tais como Ferro, Enxofre, Magnésio e Manganês em culturas de cevada, trigo, milho e girassol usando espectros de refletância, absorbância e transmitância, chegando a conclusão que a deficiência desses minerais afeta a concentração de clorofila da folhas.

Ferreira (2013) avaliou a possibilidade de utilização das técnicas NIRS, principalmente, e MIRS, associadas à quimiometria, para previsão de parâmetros de qualidade da soja brasileira e quinoa da América do Sul. As amostras de soja provenientes do Paraná foram analisadas pelas duas técnicas para previsão da composição centesimal. Os erros relativos (E%) foram baixos tanto para NIRS como para MIRS, sendo que os resultados sugerem o uso da NIRS para previsão de lipídios (0,2 a 9,2%) e o uso da MIRS para proteínas (0,2 a 5,6%), cinzas (0 a 5,0%) e umidade (0,1 a 2,0%). Os melhores modelos de calibração encontrados para soja foram para proteína e umidade, com melhores coeficientes de determinação e raiz quadrada do erro médio quadrático de calibração ($R^2 = 0,81$, RMSEC = 0,58% e $R^2 = 0,80$, RMSEC = 0,28%, respectivamente). A técnica mostrou capacidade adequada de predição para os parâmetros, incluindo lipídios, cinzas, carboidratos e fibras.

Zhai *et al.* (2013) compararam dois métodos de regressão, o SVMR (*Support Vector Machine Regression*) e o PLSR (*Partial Least Squares Regression*), com a finalidade de estimar valores de nitrogênio, fósforo e potássio presentes em 94 amostras de folhas de diversas plantas, entre elas soja, arroz, milho e gergelim, usando da técnica VIS-NIR (Espectroscopia do Infravermelho Próximo e Visível). Os melhores valores de coeficientes foram obtidos utilizando SVMR, tendo sido 0,70, 0,72 e 0,70 para N, P e K, respectivamente.

Ferreira *et al.* (2014) compararam as duas técnicas de espectroscopia NIRS e MIRS para determinar parâmetros de qualidade, como umidade, proteína, lipídios e quantidade de cinzas a partir de 40 amostras de grãos de soja da região de Ponta Grossa e Londrina. As técnicas de espectroscopia foram utilizadas juntamente com métodos de calibração baseado em algoritmo PLS. O coeficiente de determinação obtido para umidade, proteína, lipídios e quantidade de cinzas foi, respectivamente, 0,72, 0,73, 0,88 e 0,81 para técnica NIRS e 0,63, 0,87, 0,91 e 0,67 para técnica MIRS. Os resultados mostraram que ambas as técnicas de espectroscopia NIRS e MIRS tem capacidade preditiva variando em função do parâmetro observado.

Lee e Choung (2011) tiveram como objetivo avaliar o potencial da técnica NIRS em diferenciar sementes de soja geneticamente modificadas de sementes que não eram

geneticamente modificadas. Para as calibrações, foram utilizadas as técnicas de PCA e Análise Discriminante por Mínimos Quadrados Parciais (PLSDA) para classificar as sementes em dois grupos: organismos geneticamente modificados (OGM) e não-OGM. A técnica PLSDA obteve 97% de precisão, permitindo concluir que a NIRS pode ser usada para identificar soja GM, evitando assim análises químicas e sensoriais demoradas, dispendiosas e laboriosas.

Observa-se nos trabalhos que o nutriente mais estimado é o Nitrogênio e portanto, pesquisas que avancem na estimativa de outros nutrientes da soja e da folha por meio de refletância espectral podem trazer uma importante contribuição para o processo de análise foliar e de grãos.

3.2 Mineração de Dados

Um problema real quando trabalha-se com dados de espectroscopia NIR é o processamento e a interpretação das informações. Como os espectros possuem um grande conjunto de dados, é preciso uma ferramenta para aplicar e retirar as informações necessárias para a análise quantitativa (FERRARINI, 2004). Tais ferramentas constituem-se de métodos matemáticos que normalmente envolvem modelos de regressão, os quais relacionam o comportamento linear ou não, de uma variável Y com outra X. A relação entre duas variáveis pode realizar previsões sobre o comportamento de algo, como a estimativa de nutrientes.

Novos métodos tem sido desenvolvidos para resolver problemas de análise de dados multivariados, os quais se inserem na área de Mineração de Dados (MD), sendo esta a principal etapa do processo de Descoberta de Conhecimento em Base de Dados (KDD - *Knowledge Discovery in Databases*) e tem como finalidade a descoberta automática de informações úteis em grandes depósitos de dados. Esta etapa é considerada o centro do processo e se preocupa em ajustar modelos ou determinar padrões a partir dos dados selecionados. É vista como uma forma de selecionar, explorar e modelar grandes conjuntos de dados para detectar padrões de comportamento e também possui a capacidade de previsão do resultado de uma observação futura (BORGES, 2006), como a previsão da estimativa de produtividade de grãos com base nas características ambientais, do solo e da planta.

A mineração de dados utiliza técnicas que são baseadas em análise estatística e Inteligência Artificial (IA), mais especificamente Aprendizagem de Máquina. (HOLSHEIMER; SIEBES, 1994). Essas técnicas, via de regra, apresentam melhor desempenho para um

elevado número de instâncias (amostras do ponto de vista agrônomo) e menor número de atributos (variáveis). Porém, as bases de dados referentes a espectroscopia na região do NIR normalmente apresentam justamente o oposto, ou seja, um grande número de atributos (de centenas a milhares de comprimentos de onda lidos) e um número relativamente pequeno de amostras (algumas centenas, normalmente não superior a 500 amostras), comprometendo o desempenho de algoritmos tradicionais de mineração de dados. Assim, a área de mineração de dados tem enfrentado desafios em termo de efetividade e eficiência devido ao aumento da dimensão dos dados a serem analisados (SUTHA; TAMILSELVI, 2015).

3.2.1 Seleção de Atributos para Redução de Dimensionalidade

Técnicas de redução de dimensionalidade tem sido desenvolvidas para serem usadas na etapa de pré-processamento de forma a preparar os dados para que a etapa de mineração dos dados ocorra de forma mais eficiente. Basicamente, a redução tem como objetivo escolher os atributos que são relevantes para a aplicação de mineração de dados, com intuito de maximizar o desempenho com o mínimo esforço de processamento (KANTARDZIC, 2003). O processo de redução de atributos deve resultar em:

- Menos dados, para que o algoritmo de mineração de dados possa aprender mais rápido;
- Maior precisão do processo de mineração de dados para que o modelo possa generalizar melhores informações;
- Resultados simples do processo de mineração de dados, de modo que facilite o entendimento e seu uso;
- Menos atributos para que a próxima fase de coleta de dados seja facilitada através desta remoção de atributos redundantes e/ou irrelevantes.

Entre as abordagens mais comumente utilizadas para redução de dimensionalidade está a Análise de Componentes Principais (PCA - *Principal Component Analysis*) (TAN; STEINBACH; KUMAR, 2006), cujo objetivo é encontrar um novo conjunto de dimensões (atributos) que capture melhor a variabilidade dos dados. E, especificamente, a primeira dimensão é escolhida para capturar tanto da variabilidade quanto possível. A segunda dimensão é ortogonal a primeira e, sujeita a esta restrição, captura tanto da variabilidade restante quanto possível, e assim por diante.

A PCA tende a identificar os padrões lineares e as variações nos dados. Assim, pode ser usada como uma técnica de descoberta de padrões. Em alguns casos, a maioria da variabilidade dos dados pode ser capturada por uma pequena fração do conjunto total de dimensões. A redução de dimensionalidade usando PCA pode resultar em dados de dimensões relativamente baixas e pode ser possível aplicar técnicas que não funcionem bem com dados de dimensionalidade alta. Considerando-se que os ruídos nos dados são (espera-se) mais fracos que os padrões, a redução da dimensionalidade pode eliminar muito desse ruído. Isto é benéfico tanto para a mineração de dados quanto para outros algoritmos de análise de dados (TAN; STEINBACH; KUMAR, 2006).

Oliveira *et al.* (2014) apresentaram um procedimento que combina métodos de busca e análise de componentes principais supervisionada e não supervisionada para a tarefa de seleção de variáveis. O procedimento teve como objetivo remover as variáveis irrelevantes e de pouca influência sobre a variação dos dados e avaliar o impacto da seleção sobre tarefas de regressão e classificação. Os testes foram realizados com base na seleção de variáveis para indução de modelos lineares multivariados e redes neurais artificiais mediante uma base de dados de agricultura de precisão. O procedimento permitiu uma solução de custo-benefício satisfatório em relação a dimensionalidade e a acurácia do modelo.

/

Mitra, Murthy e Pal (2002) desenvolveram um algoritmo não supervisionado de seleção de atributos adequado para conjunto de dados de altas dimensões. O método é baseado na medida da similaridade entre os atributos e, por não utilizar métodos de busca, o que os autores consideram a maior vantagem em relação aos outros algoritmos de seleção de atributos apresenta ganhos de performance que o deixam mais rápido. Os autores introduziram uma medida para a seleção de atributos. É também demonstrada como a redundância e a perda de informação em seleção de atributos pode ser quantificada com a medida da entropia.

Chakrabarti e Mehrotra (2000) propuseram em seu trabalho uma técnica para redução de dimensionalidade chamada Redução de Dimensionalidade Local (LDR - *Local Dimensionality Reduction*) que tenta encontrar correlações locais nos dados e realiza a redução de dimensionalidade em *clusters* (agrupamentos) de dados localmente correlacionados. Desenvolveram uma estrutura de índices que explora os *clusters* correlacionados para suportar consultas eficientemente sobre conjunto de dados de altas dimensões. Os experimentos em conjuntos de dados, tanto sintéticos como reais, mostram que a técnica

reduz a dimensionalidade dos dados com uma perda de informações significativamente menor em relação a técnicas de redução de dimensionalidade globais, por exemplo, a PCA.

Jia *et al.* (2013) desenvolveram um algoritmo de agrupamento espectral com características ponderadas com base em conhecimento de entropia (FWKE-SC - *Feature Weighted Spectral Clustering Algorithm based Knowledge Entropy*). O algoritmo utiliza o conceito de conhecimento de entropia para avaliar a importância de cada atributo, que pode ser usado como peso de atributos, e, em seguida, aplica-se o método de agrupamento espectral nos dados. Os resultados mostram que o algoritmo FWKE-SC trabalha com dados de alta dimensionalidade muito bem e tem melhor robustez e capacidade de generalização.

Kanth *et al.* (1999) propuseram uma nova técnica para redução de dimensionalidade baseada em SVD (Decomposição de valor singular - *Singular Value Decomposition*) para bases de dados dinâmicas. Quando a distribuição de dados muda consideravelmente de modo a degradar a precisão de consulta, pode-se recalcular a SVD e incorporá-lo na estrutura de índice existente. Para recalcular a SVD, os autores propuseram uma nova técnica que utiliza dados agregados do índice existente, em vez de dados inteiros. Esta técnica reduz o tempo de SVD sem comprometer a precisão de consulta. Em seguida, analisou-se eficientes maneiras de incorporar o cálculo da SVD transformada na estrutura de índice existente sem degradar o tempo de resposta das consultas subsequentes. Estas técnicas reduziram o tempo de computação por um fator de 20 em experimentos na cor e textura de vetores de imagem.

Kovalenko, Rippke e Hurburgh (2006) buscaram em seu trabalho determinar equações de calibração com intuito de estimar a composição de aminoácidos em grãos de soja utilizando regressão de mínimos quadrados parciais (PLS), redes neurais artificiais (RNA) máquina de vetor de suporte (SVM) para cinco modelos de base de dados de espectroscopia do infravermelho próximo (NIRS). A validação dos modelos de previsão resultaram em valores R^2 de 0,04 para triptofano e 0,91 para leucina e lisina. No trabalho, a capacidade preditiva das calibrações estavam dependentes das respectivas correlações de aminoácidos com a proteína de referência. O desempenho do método PLS e da SVM foi significativamente melhor do que a RNA.

Uma ferramenta de seleção de atributos recentemente desenvolvida é o *SynthesisFS* (SANTOS, 2016), esse software utiliza separadamente cada técnica de seleção de atributos implementada no ambiente Weka e criar um ranqueamento de atributos com

base na síntese dos resultados de execução das diversas técnicas. Neste processo de ranqueamento desenvolvido no *SynthesisFS*, cada atributo recebe uma pontuação, sendo que quanto maior esse valor, maior a importância do atributo em relação ao meta. Para cada técnica utilizada, os atributos selecionados recebem uma determinada pontuação que varia de acordo com algumas condições. Uma dessas condições é o método de busca empregado, que no caso do *Ranker*, os atributos selecionados possuem uma ordem em relação a correlação com atributo meta. Caso o método de seleção seja feito através do modo de treinamento completo os atributos são considerados a partir de suas quantidades pertencentes ao subconjunto selecionado. Se o modo for validação cruzada considera-se a porcentagem de *folds* de cada atributo selecionado no método, ou a quantidade de atributos selecionados em ao menos um *fold*. Por fim, a pontuação final de cada atributo é obtida através da soma de sua pontuação para cada método de seleção utilizado, resultando em um escore que define um *ranking* para os atributos selecionados. Após os atributos estarem ranqueados, a escolha dos atributos é feita a partir daquele que obteve o maior valor até um limite de 10 % e 20 % do primeiro atributo selecionado. São gerados dois subconjuntos de atributos, o primeiro constituído de atributos que possuam pontuações até 10 % do primeiro e outro com atributos que possuam pontuações com valores até 20 % do primeiro. Esse método de síntese de seleção de atributos se mostra promissor no processo de redução de dimensionalidade.

3.3 Algoritmos para Tarefa de Regressão

Dentre vários algoritmos de mineração de dados para geração de modelos de regressão, três serão destacados nesse trabalho, sendo eles: o M5Rules, SMOREg e LinearRegression.

O algoritmo M5Rules, apresentado por (HOLMES; HALL; PRANK, 1999), cria modelos de regras de produção baseadas em árvores que são geradas repetidamente e a melhor regra é selecionada a cada interação. O processo para gerar as regras a partir de modelos de árvores funciona de modo que um modelo de árvores é aplicado ao conjunto de dados com treinamento completo e uma árvore gerada é podada, a melhor folha desta é transformada em regra e a árvore restante é descartada. Todas as demais árvores geradas cobertas pela regra são removidas do conjunto de dados, sendo este processo aplicado de forma recursiva às instâncias restantes e completo quando todas são cobertas pela regra. Este algoritmo foi utilizado por Teixeira *et al.* (2014), dentre outros, afim de encontrar correlações entre dados espectrais Vis-NIR afim de prever argila e MO em solos;

O algoritmo SMOReg disponível na ferramenta WEKA, é uma implementação do algoritmo SMO proposto por Platt (1999) e melhorado por Shevade *et al.* (2000) que propõem um algoritmo iterativo que cria modelos de regressão utilizando-se da SVM (VAPNIK, 2013). O algoritmo SMO tem como intuito diminuir a quantidade de operações aritméticas e consequentemente diminuir o tempo de processamento da SVM.

Dentre os algoritmos que codificam o SMV, o SMO possui uma complexidade computacional menor em relação a outros por não trabalhar com inversão de uma matriz de kernel. Esta matriz quadrada, costuma ter tamanho proporcional ao número de amostras que constituem os vetores de suporte (KINTO, 2011). Este algoritmo é de especial interesse por ser considerado um método adequado para dados de alta dimensionalidade, utilizado por Kovalenko, Rippke e Hurburgh (2006) e Zhai *et al.* (2013).

Finalmente, o algoritmo LinearRegression descrito por Witten *et al.* (1999) faz uso da regressão linear que é uma técnica estatística que correlaciona uma variável com outra. Quando a função f que relaciona duas variáveis é do tipo $f(X) = a + bX$ temos um modelo de regressão simples. A variável X é a variável independente da equação, enquanto $Y = f(X)$ é a variável dependente das variações de X .

O modelo de regressão é chamado de multivariado quando envolve uma relação com mais de duas variáveis. Isto é, quando o comportamento de Y é explicado por mais de uma variável independente, no caso $X_1, X_2, \dots, X_n$ (TAN; STEINBACH; KUMAR, 2006). Ao final gera uma função de saída, um modelo com as variáveis, visando prever ou estimar outra (TAN; STEINBACH; KUMAR, 2006). Equações geradas tem função preditiva e uma ampla aplicabilidade, como também na agricultura (TAN; STEINBACH; KUMAR, 2006; WITTEN *et al.*, 2016).

4 MATERIAL E MÉTODOS

4.1 Área de Coleta de Amostras de Folha e Grãos de Soja

Este estudo foi realizado através de amostras coletadas em seis áreas experimentais localizadas no Estado do Paraná na região dos Campos Gerais, sendo elas: Fazendas Estância II, Regina, Massinham, Eumir, Santa Cruz e Fazenda Escola da UEPG, compreendendo 231 amostras de folhas de soja e 285 de grãos de soja. Tais amostras bem como os resultados das análises laboratoriais realizados conforme métodos descritos em (MALAVOLTA; VITTI; OLIVEIRA, 1989) foram ambos fornecidos pelo Laboratório de Fertilidade do Solo da UEPG (Universidade Estadual de Ponta Grossa), tendo sido coletadas durante o ano agrícola de 2013-2014. O Nitrogênio (N) foi extraído através de digestão sulfúrica e determinado pelo método semi-micro-Kjeldahl. As análises de Fósforo (P) por meio de digestão nítrico-perclórica e determinação por espectroscopia de absorção molecular, Potássio (K) por fotometria de chama, Enxofre (S) através de turbidimetria do sulfato de bário, Cálcio (Ca), Magnésio (Mg), Cobre (Cu), Ferro (Fe), Zinco (Zn) e Manganês (Mn) por espectroscopia de absorção atômica e o Boro (B) foi extraído por incineração e determinado por espectrofotometria de absorção molecular, pelo método da azometina H.

De acordo com o sistema de classificação climática de *Köppen* baseado na vegetação, temperatura e pluviosidade, a região dos Campos Gerais pertence ao clima Cfb - Clima temperado; temperatura média no mês mais frio abaixo de 18°C, com verões frescos, temperatura média no mês mais quente abaixo de 22°C e sem estação seca definida, a precipitação média anual é de 1422 milímetros, sendo janeiro o mês mais chuvoso e agosto o mais seco (IAPAR, 2015).

4.2 Preparo e Análise - Equipamento NIRS e MIRS

Os dois conjuntos de amostras, tanto de folhas como de grãos, foram analisados a partir de um espectrofotômetro NIR da marca FOSS, modelo DA 1650, como mostra a Figura 1, instalado no Centro Interdisciplinar de Pesquisa e Pós-Graduação (CIPP) da UEPG.

Para aquisição dos espectros o software utilizado foi o FOSS Mosaic Solo versão 6.6, que trabalha com intervalos de comprimento de onda entre 1.100 nm e 1.650 nm com 2 nm entre as varreduras.

O Software WinISI II, versão 4.6 foi utilizado para a montagem das curvas e ex-

tração dos dados em formato txt. Esses espectros foram incorporados ao banco de dados, no qual para cada amostra, foi armazenada sua correspondente curva espectral.



Figura 1: Espectrofotômetro NIR, marca FOSS, modelo DA 1650, onde foram lidas as amostras de folhas e grãos.

A Figura 2, apresenta uma amostra de folha de soja dentro de um recipiente próprio, denominado SmallCup, para uso no espectrofotômetro NIR, sendo que a quantidade de amostra de soja inserida nesse acessório foi de 40 ml.



Figura 2: Amostra com 40 ml de folha de soja no acessório SmallCup do espectrofotômetro NIR.

O conjunto de amostras de folhas de soja foram também analisados no espectrofotômetro da marca Shimadzu, modelo IRPrestige-21 (Figura 3), localizado no CIPP da UEPG, que possui comprimentos de onda na região do infravermelho médio, sendo que seus resultados são comparados aos do NIRS com objetivo de analisar e concluir quais técnicas obtiveram melhores valores para estimativa de teores de nutrientes. O software utilizado foi o Shimadzu IRsolution, versão 1.6 para obtenção dos espectros. Estes espectros foram obtidos a partir da varredura efetuada no equipamento na faixa de números de onda de 4.000 a 400 cm^{-1} , sendo os valores correspondentes em comprimentos de onda de 2.500 a 25.000 nm. A criação das curvas e extração dos dados em formato txt foram

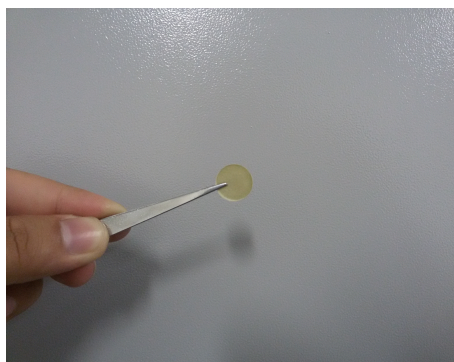
efetuadas através do software IRsolution.



Figura 3: Espectrofotômetro MIR, marca Shimadzu, modelo IRPrestige-21, utilizado para leituras de amostras de folhas.

As amostras de folhas de soja precisaram ser preparadas antes de serem analisadas no equipamento MIR, no qual foi utilizado a técnica de pastilhas com brometo de potássio (KBr). Para cada amostra de folha de soja, foi criada uma pastilha (Figura 4(a)), utilizando as quantidades de 0,1 g de KBr e 0,001 g de amostra de folha para cada pastilha.

O KBr e as amostra de folhas de soja foram primeiramente pesadas utilizando uma balança analítica de precisão disponibilizada pelo CIPP da UEPG. Logo após serem pesadas, foram prensadas por 5 min em um pastilhador da Shimadzu (Figura 4(b)). Depois de feitas, as pastilhas foram colocadas em um acessório próprio do MIR e então realizada as análises.



(a) Pastilha feita da amostra de folha de soja com KBr.



(b) Pastilhador da marca Shimadzu.

Figura 4: Descrição de pastilha feita da amostra de folha de soja com KBr (Figura 4(a)) e Pastilhador da marca Shimadzu (Figura 4(b)).

4.3 Elaboração da Base de Dados

A base de dados desse trabalho consistiu em um conjunto de dados foliares (DF) e um conjunto de dados de grãos (DG) de soja. Tanto para o conjunto DF como para o DG, foram armazenados os teores de N, P, K, Ca, Mg, S, B, Cu, Fe, Zn e Mn, os quais foram obtidos a partir de análises realizadas em laboratório, respectivamente para folhas e para grãos, conforme a Tabela 2.

Tabela 2: Descrição de teores de nutrientes obtidas a partir das análises de laboratório para folhas e grãos de soja, como também os valores mínimos, máximos e média.

Nutrientes	Unidade	Interv. de valores p/ Folhas			Interv. de valores p/ Grãos		
		Mín	Máx	Média	Mín	Máx	Média
Nitrogênio	g/kg	5,068	60,76	47,87	29,96	88,48	57,54
Fósforo	g/kg	1,36	97,60	11,45	2,55	45,30	9,90
Potássio	g/kg	0	40,81	22,45	0,48	32,69	18,76
Cálcio	g/kg	0,17	11,53	8,04	0,91	4,53	2,19
Magnésio	g/kg	0,017	7,92	3,28	1,26	5,88	2,74
Enxofre	g/kg	1,03	96,6	8,73	0,99	40,90	8,18
Cobre	mg/kg	0,01	42,70	15,17	2,22	17,82	11,58
Manganês	mg/kg	17,74	246,39	70,10	10,89	846,30	57,98
Ferro	mg/kg	3,91	384,77	119,57	0,59	673,54	83,35
Zinco	mg/kg	10,93	202,24	75,23	13,08	62,35	31,90
Boro	mg/kg	33,85	79,11	52,98	34,17	64,45	44,05

Para criação dessa base de dados, foram utilizadas as 231 amostras de folhas de soja e 285 de grãos de soja. As amostras de grãos de soja foram analisadas no espectrofotômetro NIR e as de folhas foram analisadas nos espectrofotômetros NIR e MIR.

Ao final das análises, três bases de dados foram geradas, duas com dados de análises foliares NIR e MIR, e uma com dados de análises de grãos de soja NIR. Antes da etapa de seleção de atributos as três bases sofreram transformações matemáticas em seus dados de espectros, no qual as funções quadrática, neperiana e inversa foram utilizadas com intuito de aumentar a diferença entre os valores dos espectros. As bases de dados denominadas transformadas receberam os três valores transformados pelos métodos, sendo excluídos os valores iniciais. Logo, a quantidade de atributos passou a ser três vezes mais em comparação as bases de dados não transformadas.

Após as transformações as bases foram utilizadas para estimativa de teores dos nutrientes, sendo três bases originais, denominadas não transformadas e três bases transformadas.

4.4 Seleção de Atributos

Após a criação das bases de dados, foi realizado um pré-processamento aplicando métodos computacionais de seleção de atributos utilizando a ferramenta *SynthesisFS*, que permite realizar seleção dos atributos fazendo uso das técnicas disponíveis pela ferramenta Weka, conforme descrito na seção 3.2.1.

4.5 Mineração de Dados – Determinação de Modelos de Regressão

Uma vez efetuada a seleção de atributos, o próximo passo foi desenvolver modelos de regressão para estimativa de teor de nutrientes nas folhas e nos grãos de soja. Foram usados como medida de avaliação da qualidade dos modelos o erro quadrático (RRSE) e o coeficiente de correlação, sendo para este estipulado um limiar de 0,70 a ser atingido pelos algoritmos. Os métodos de regressão utilizados foram os algoritmos M5Rules, LinearRegression e SMOReg, sendo este indicado para dados de alta dimensionalidade. Assim, é possível avaliar o efeito do uso de técnicas de seleção de atributos comparadas a resultados gerados sem o emprego prévio dessas técnicas. A avaliação dos modelos de regressão gerados pelos algoritmos foram feitas por meio do coeficiente de correlação e erro relativo.

5 RESULTADOS

Nesta seção são apresentados os resultados obtidos pelas análises de folhas e grãos, realizadas nos espectrofotômetros NIR e MIR, detalhando separadamente cada nutriente e as técnicas utilizadas para geração do modelo de regressão. Três algoritmos foram utilizados para geração dos modelos, a SMOReg, recomendada para bases de dados de alta dimensionalidade e por gerar equações de regressão; algoritmo M5Rules baseado em árvores de decisão com funções de regressão e tem como objetivo gerar um modelo de previsão sob a forma de lista de regras; e o algoritmo LinearRegression, todos disponíveis na ferramenta WEKA v. 3.6.13.

5.1 Análise Foliar

Nesta seção é apresentado os resultados referentes a análise foliar a partir das bases de dados geradas nos espectrofotômetros NIR e MIR a partir da utilização de três técnicas de análise de regressão, a M5Rules, SMOReg e a LinearRegression.

5.1.1 Espectroscopia NIRS

A base de dados analisada nesta seção contém dados de 200 amostras de folhas de soja analisadas no equipamento NIR modelo DA 1650 e 275 atributos correspondendo aos comprimentos de onda de 1.100 até 1.648 nm referentes ao espectro do infravermelho próximo, sendo a leitura realizada a cada 2 nm. As Tabelas 3, 4 e 5 apresentam uma síntese dos resultados utilizados, respectivamente, 10%, 20% e todos os atributos.

Tabela 3: Resultados da análise foliar utilizando redução da dimensionalidade em 10 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	18	0,66	77,14	0,48	87,17	0,68	72,64
P	6	0,05	110,22	0,36	93,93	0,69	72,35
K	16	0,50	90,65	0,73	69,17	0,84	54,39
Ca	5	0,21	97,17	0,23	96,91	0,24	96,86
Mg	5	0,54	84,11	0,51	85,86	0,65	75,39
S	15	0,60	100,20	0,55	122,21	0,15	91,00
Cu	4	0,55	86,50	0,53	84,20	0,60	79,71
Mn	6	0,52	88,94	0,52	85,15	0,58	81,16
Fe	3	0,31	95,72	0,44	90,99	0,62	77,98
Zn	4	0,33	99,36	0,63	77,52	0,50	86,30
B	28	0,46	88,38	0,45	88,72	0,44	89,18

Tabela 4: Resultados da análise foliar utilizando redução da dimensionalidade em 20 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	32	0,67	75,82	0,53	84,58	0,64	77,71
P	28	0,71	102,75	0,83	55,24	0,92	38,63
K	36	0,63	82,88	0,73	68,08	0,92	39,78
Ca	15	0,22	96,88	0,29	95,18	0,46	88,69
Mg	8	0,61	78,68	0,61	79,20	0,67	73,84
S	31	0,75	95,17	0,76	64,95	0,88	47,88
Cu	9	0,59	82,21	0,58	81,15	0,68	72,78
Mn	8	0,54	87,38	0,53	84,31	0,68	72,83
Fe	14	0,50	86,27	0,58	81,02	0,74	67,23
Zn	15	0,33	99,06	0,55	83,61	0,75	65,55
B	46	0,47	87,98	0,46	88,26	0,41	90,67

Tabela 5: Resultados da análise foliar sem utilizar redução da dimensionalidade, totalizando 275 atributos.

Nutriente	Algoritmo					
	SMOReg		M5Rules		LinearRegression	
	r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	0,70	70,85	0,49	86,72	0,49	86,36
P	0,76	44,19	0,79	61,27	0,61	79,32
K	0,78	62,85	0,78	62,64	0,56	82,22
Ca	0,36	92,40	0,12	99,38	0,16	98,49
Mg	0,70	71,44	0,63	77,86	0,63	77,30
S	0,76	70	0,75	65,84	0,57	81,63
Cu	0,70	71,72	0,59	80,77	0,60	79,65
Mn	0,74	67,56	0,72	69,24	0,56	82,19
Fe	0,67	75,13	0,43	92,13	0,48	87,13
Zn	0,66	77,86	0,57	81,70	0,39	91,44
B	0,51	85,27	0,43	89,83	0,44	89,48

As subseções abaixo detalham os resultados apresentados.

5.1.1.1 N - Nitrogênio

A partir da base de dados de folhas geradas através da seleção de atributos foram obtidos os seguintes resultados descritos na Tabela 6.

O algoritmo SMOReg foi o que obteve o maior coeficiente de correlação, $r = 0,70$, utilizando a quantidade total de atributos. Levando em consideração um limiar de correlação de 0,70 é válido destacar o algoritmo LinearRegression que obteve uma correlação de 0,68, uma diferença de 0,02 para o SMOReg, mas no caso utilizou 18 atributos ao invés da quantidade total.

Tabela 6: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de N com diferentes quantidades de atributos.

	Quantidades de Atributos					
	18		32		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,66	77,14 %	0,67	75,82 %	0,70	70,85 %
M5Rules	0,48	87,17 %	0,53	84,58 %	0,49	86,72 %
LinearRegression	0,68	72,64 %	0,64	77,71 %	0,49	86,36 %

5.1.1.2 P - Fósforo

Os três algoritmos alcançaram resultados de correlação acima de 0,70 utilizando 28 atributos dos 275 totais disponíveis na base de dados, sendo que a LinearRegression obteve a melhor correlação $r = 0,92$, e também o menor valor do erro quadrático, $RRSE = 38,63 \%$. A Tabela 7 apresenta os seguintes resultados para o nutriente P. No Quadro 1 descrito os 28 atributos utilizados pelos algoritmos.

Tabela 7: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.

	Quantidades de Atributos					
	6		28		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,05	110,22 %	0,71	102,75 %	0,76	44,19 %
M5Rules	0,36	93,93 %	0,83	55,24 %	0,79	61,27 %
LinearRegression	0,69	72,35 %	0,92	38,63 %	0,61	79,32 %

Quadro 1: Descrição dos 28 atributos utilizados na execução dos algoritmos gerando correlações acima de 0,70 no intervalo de 1.128 a 1.568 nm.

Comprimentos de Onda
1.118 - 1.124 - 1.184 - 1.196 - 1.204 - 1.216 - 1.238
1.246 - 1.264 - 1.276 - 1.290 - 1.302 - 1.310 - 1.382
1.404 - 1.410 - 1.436 - 1.450 - 1.456 - 1.476 - 1.482
1.492 - 1.500 - 1.512 - 1.538 - 1.550 - 1.558 - 1.568

O algoritmo LinearRegression destacou-se ao obter uma correlação de 0,69 utilizando-se de apenas 6 atributos, como demonstra o Quadro 2 utilizando comprimentos de onda com amplitude de aproximadamente 100 nm (1.404 a 1.500 nm.)

Quadro 2: Descrição dos 6 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,69 no intervalo de 1.404 a 1.500 nm.

Comprimentos de Onda
1.404 - 1.410 - 1.450 - 1.456 - 1.482 - 1.500

5.1.1.3 K - Potássio

A Tabela 8 apresenta os resultados dos coeficientes de correlação e os erros relativos de raiz quadrada obtidos a partir da aplicação dos algoritmos de regressão em relação ao K.

Tabela 8: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.

	Quantidades de Atributos					
	16		36		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,50	90,65 %	0,63	82,88 %	0,78	62,85 %
M5Rules	0,73	69,17 %	0,73	68,08 %	0,78	62,64 %
LinearRegression	0,84	54,39 %	0,92	39,78 %	0,56	82,22 %

Ambos algoritmos de regressão obtiveram resultados de coeficientes de correlação acima de 0,70, tendo um maior destaque para a LinearRegression que obteve uma correlação de 0,92 utilizando 36 dos 275 atributos da base de folhas e uma correlação de 0,84 utilizando 16 atributos para geração do modelo de regressão. O algoritmo M5Rules que gera regras a partir de uma árvore de decisão obteve um resultado acima do limiar estipulado de 0,70, que foi de 0,73 utilizando 16 atributos, este algoritmo esteve acima do limiar em seus três modos de execução, tanto para 16, 36 ou 275 atributos.

O Quadro 3 apresenta os 36 atributos utilizados pelos dois algoritmos que obtiveram valores de correlação acima de 0,70, o M5Rules e a LinearRegression, e contém 16 atributos que foram utilizados pelos dois algoritmos, no qual obtiveram correlações de 0,73 e 0,84 respectivamente.

Quadro 3: Descrição dos 36 atributos utilizados na execução dos algoritmos LinearRegression e M5Rules gerando correlações acima de 0,70 e também 16 atributos pelo qual os dois algoritmos obtiveram correlações de 0,73 e 0,84 respectivamente no intervalo de 1.104 a 1.642 nm.

Comprimentos de Onda
1.104 - 1.112 - 1.118 - 1.124 - 1.132 - 1.138 - 1.154 - 1.170
1.182 - 1.190 - 1.196 - 1.208 - 1.214 - 1.228 - 1.236
1.242 - 1.248 - 1.262 - 1.284 - 1.302 - 1.316 - 1.336
1.352 - 1.384 - 1.402 - 1.424 - 1.430 - 1.442 - 1.456
1.488 - 1.512 - 1.550 - 1.566 - 1.596 - 1.634 - 1.642

5.1.1.4 Ca - Cálcio

Segue abaixo na Tabela 9 os resultados obtidos através da aplicação dos algoritmos para estimativa de Ca a partir da base de dados foliares de soja contendo 275 atributos

correspondentes aos comprimentos de onda do espectro NIR.

Tabela 9: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.

	Quantidades de Atributos					
	5		15		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,21	97,17 %	0,22	96,88 %	0,36	92,40 %
M5Rules	0,23	96,91 %	0,29	95,18 %	0,12	99,38 %
LinearRegression	0,24	96,86 %	0,46	88,69 %	0,16	98,49 %

Os algoritmos utilizados para geração de modelos não obtiveram resultados acima do limiar estipulado de 0,70, sendo que o maior resultado atingido para o nutriente Ca foi de $r = 0,46$ pelo algoritmo LinearRegression utilizando-se de 15 atributos.

5.1.1.5 Mg - Magnésio

Os algoritmos utilizados para geração de modelos de regressão, apenas a SMO-reg atingiu o resultado acima do limiar, obtendo um coeficiente de correlação $r = 0,70$ utilizando-se do conjunto completo de atributos. A Tabela 10 apresenta os resultados obtidos para nutriente Mg.

Tabela 10: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.

	Quantidades de Atributos					
	5		8		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,54	84,11 %	0,61	78,68 %	0,70	71,44 %
M5Rules	0,51	85,86 %	0,61	79,20 %	0,63	77,86 %
LinearRegression	0,65	75,39 %	0,67	73,84 %	0,63	77,30 %

Em apenas um dos casos o limiar de correlação 0,70 foi atingido, na situação em que o algoritmo SMOReg obteve um coeficiente $r = 0,70$ utilizando-se de todos os atributos da base.

5.1.1.6 S - Enxofre

Com atributo meta S os resultados de correlações foram acima de 0,70, sendo que a LinearRegression destacou-se por obter um coeficiente de correlação de 0,88 utilizando 31 atributos. A Tabela 11 apresenta os resultados obtidos pelos algoritmos de regressão.

Tabela 11: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.

	Quantidades de Atributos					
	15		31		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,60	100,20 %	0,75	95,17 %	0,76	70 %
M5Rules	0,55	122,21 %	0,76	64,95 %	0,75	65,84 %
LinearRegression	0,15	91 %	0,88	47,88 %	0,57	81,63 %

Os algoritmos obtiveram correlações acima de 0,70 utilizando-se de 31 atributos, sendo que a LinearRegression destacou-se por obter o maior valor de correlação. O Quadro 4 apresenta os 31 atributos que foram utilizados pelos algoritmos.

Quadro 4: Descrição dos 31 atributos utilizados na execução dos algoritmos gerando correlações acima de 0,70 no intervalo de 1.102 a 1.600 nm.

Comprimentos de Onda
1.102 - 1.114 - 1.124 - 1.184 - 1.192 - 1.200 - 1.226
1.256 - 1.286 - 1.294 - 1.308 - 1.316 - 1.370
1.402 - 1.416 - 1.424 - 1.440 - 1.446 - 1.458
1.466 - 1.474 - 1.480 - 1.494 - 1.510 - 1.516
1.522 - 1.538 - 1.544 - 1.550 - 1.588 - 1.600

5.1.1.7 Cu - Cobre

Para o nutriente Cu, apenas o algoritmo SMOReg atingiu o limiar de 0,70 para coeficientes de correlação, utilizando-se dos 275 atributos disponíveis. O algoritmo LinearRegression foi o que mais aproximou-se do limiar, obtendo um valor de 0,68 de correlação, neste caso utilizando 9 atributos. Na Tabela 12 é possível observar os valores de correlações obtidos de cada algoritmo.

Tabela 12: Resultados de correlações obtidos pelos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.

	Quantidades de Atributos					
	4		9		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,55	86,50 %	0,59	82,21 %	0,70	71,72 %
M5Rules	0,53	84,20 %	0,58	81,15 %	0,59	80,77 %
LinearRegression	0,60	79,71 %	0,68	72,78 %	0,60	79,65 %

O Quadro 5 apresenta os 9 atributos utilizados pela LinearRegression, que obteve correlação $r = 0,68$, próxima do limiar de 0,70, mas com uma quantidade menor de atributos em comparação a SMOReg que obteve $r = 0,70$ utilizando a totalidade dos atributos.

Quadro 5: Descrição dos 9 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,68 no intervalo de 1.100 a 1.182 nm.

Comprimentos de Onda
1.100 - 1.106 - 1.124 - 1.130 - 1.140 1.154 - 1.162 - 1.172 - 1.182

5.1.1.8 Mn - Manganês

Para o Mn, dois algoritmos obtiveram coeficientes de correlação acima de 0,70, a SMOReg e o M5Rules, com 0,74 e 0,72 respectivamente, mas utilizando de 275 atributos. O algoritmo LinearRegression obteve um resultado próximo ao limiar de 0,70, que foi de 0,68, mas neste caso utilizou-se de 8 atributos. A Tabela 13 apresenta os coeficientes de correlação obtidos pelos algoritmos.

Tabela 13: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.

	Quantidades de Atributos					
	6		8		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,52	88,94 %	0,54	87,38 %	0,74	67,56 %
M5Rules	0,52	85,15 %	0,53	84,31 %	0,72	69,24 %
LinearRegression	0,58	81,16 %	0,68	72,83 %	0,56	82,19 %

O resultado obtido pela LinearRegression mostrou-se interessante quando comparado a SMOReg, que utilizando-se de uma quantidade menor de atributos o resultado aproximou-se do limiar de 0,70, sendo a diferença entre os algoritmos de 0,02.

5.1.1.9 Fe - Ferro

Para o Fe como atributo meta, em apenas um dos casos foi obtido um coeficiente de correlação acima de 0,70, sendo o algoritmo LinearRegression que atingiu $r = 0,74$ utilizando-se de 14 atributos. A SMOReg obteve um coeficiente de 0,67, aproximando-se do limiar, mas no caso utilizou todo o conjunto de atributos. A Tabela 14 apresenta os resultados obtidos pelos algoritmos.

No Quadro 6 é apresentado os 14 atributos utilizados pela LinearRegression que obteve um coeficiente de correlação $r = 0,74$. A LinearRegression demonstrou que mesmo sendo um algoritmo menos robusto que a SMOReg, nos casos em que são utilizados uma quantidade menor de atributos, ela obteve um resultado melhor.

Tabela 14: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Fe com diferentes quantidades de atributos.

	Quantidades de Atributos					
	3		14		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,31	95,72 %	0,50	86,27 %	0,67	75,13 %
M5Rules	0,44	90,99 %	0,58	81,02 %	0,43	92,13 %
LinearRegression	0,62	77,98 %	0,74	67,23 %	0,48	87,13 %

Quadro 6: Descrição dos 14 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,74 no intervalo de 1.178 a 1.638 nm.

Comprimentos de Onda
1.178 - 1.200 - 1.214 - 1.226 - 1.238 - 1.338 - 1.366
1.374 - 1.432 - 1.504 - 1.512 - 1.546 - 1.596 - 1.638

5.1.1.10 Zn - Zinco

Para o Zn, um algoritmo obteve coeficiente de correlação acima do limiar, no caso a LinearRegression, que obteve $r = 0,75$ utilizando 15 atributos. A Tabela 15 apresenta os resultados obtidos após a execução dos algoritmos.

Tabela 15: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.

	Quantidades de Atributos					
	4		15		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,33	99,36 %	0,33	99,06 %	0,66	77,86 %
M5Rules	0,63	77,52 %	0,55	83,61 %	0,57	81,70 %
LinearRegression	0,50	86,30 %	0,75	65,55 %	0,39	91,44 %

O Quadro 7 apresenta os 15 atributos utilizados pela LinearRegression para obter um coeficiente de correlação de 0,75.

Quadro 7: Descrição dos 15 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,75 no intervalo de 1.100 a 1.566 nm.

Comprimentos de Onda
1.100 - 1.106 - 1.122 - 1.132 - 1.150 - 1.158 - 1.212 - 1.276
1.282 - 1.298 - 1.354 - 1.366 - 1.372 - 1.560 - 1.566

5.1.1.11 B - Boro

Para o B, os algoritmos não obtiveram resultados acima do limiar de 0,70, sendo que o algoritmo SMOReg, foi o que atingiu o maior resultado obtendo um coeficiente

de correlação $r = 0,51$ utilizando os 275 atributos. A Tabela 16 apresenta os resultados obtidos pelos algoritmos.

Tabela 16: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.

Algoritmo	Quantidades de Atributos					
	28		46		275	
	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,46	88,38 %	0,47	87,98 %	0,51	85,27 %
M5Rules	0,45	88,72 %	0,46	88,26 %	0,43	89,83 %
LinearRegression	0,44	89,18 %	0,41	90,67 %	0,44	89,48 %

5.1.2 Espectroscopia MIRS

Nesta seção são apresentados os resultados obtidos a partir da análise de uma base de dados que contém 231 amostras de folhas de soja analisadas no equipamento F-TIR (*Fourier Transform Infrared Spectroscopy*) modelo IRPrestige-21, que possui números de onda do 4.000 cm^{-1} até 400 cm^{-1} e uma quantidade de 1.868 atributos compreendendo a região do infravermelho médio. As Tabelas 17, 18 e 19 apresentam uma síntese dos resultados utilizados, respectivamente, 10%, 20% e todos os atributos.

Tabela 17: Resultados da análise foliar utilizando redução da dimensionalidade em 10 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	10	0,50	87,47	0,34	94,20	0,57	81,92
P	3	0,36	107,16	0,27	95,88	0,36	92,86
K	3	0,32	102,32	0,26	96,64	0,40	91,55
Ca	5	0,08	99,28	0,07	99,57	0,0007	101,58
Mg	9	0,46	90,82	0,42	93,83	0,53	84,64
S	4	0,32	108,70	0,23	96,90	0,29	95,35
Cu	3	0,24	102	0,29	95,51	0,46	88,18
Mn	2	0,47	91,86	0,29	95,20	0,46	88,10
Fe	1	0,15	112,76	0,005	100,79	0,03	99,90
Zn	6	0,50	90,16	0,30	94,83	0,51	85,34
B	4	0,25	98,84	0,15	98,88	0,11	99,32

Tabela 18: Resultados da análise foliar utilizando redução da dimensionalidade em 20 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	22	0,52	86,50	0,34	95,11	0,59	81,61
P	7	0,63	84,34	0,57	81,89	0,39	92,22
K	8	0,53	86,22	0,52	85,14	0,46	88,34
Ca	8	0,21	97,38	0,05	99,75	0,40	91,35
Mg	27	0,49	90,07	0,42	91,20	0,54	85,35
S	7	0,31	108,82	0,23	97,87	0,31	95,03
Cu	6	0,37	95,50	0,33	93,92	0,46	88,15
Mn	4	0,50	89	0,27	95,99	0,53	84,21
Fe	3	0,43	91,37	0,31	95,25	0,42	90,21
Zn	23	0,50	89,60	0,37	92,72	0,74	67,31
B	37	0,30	99,53	0,25	97,03	0,39	110,42

Tabela 19: Resultados da análise foliar sem utilizar redução da dimensionalidade, totalizando 1868 atributos.

Nutriente	Algoritmo					
	SMOReg		M5Rules		LinearRegression	
	r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	0,58	83,05	0,40	93,61	0,41	89,84
P	0,76	65,97	0,54	85,28	0,47	88,32
K	0,76	64,54	0,50	90,92	0,40	95,45
Ca	0,41	90,90	0,26	101,60	0,36	88,34
Mg	0,65	75,90	0,47	88,04	0,50	90,05
S	0,74	68,71	0,37	99,38	0,30	96,46
Cu	0,85	53,07	0,64	76,58	0,38	92,12
Mn	0,77	63,42	0,52	86,61	0,49	82,52
Fe	0,81	58,45	0,46	89,25	0,39	92,12
Zn	0,84	53,97	0,60	79,46	0,50	89,39
B	0,67	74,34	0,41	92,81	0,42	112,79

As subseções abaixo detalham os resultados apresentados.

5.1.2.1 N - Nitrogênio

Nenhum dos algoritmos obteve resultados que atingissem o limiar de 0,70, sendo que o algoritmo que se aproximou desse resultado foi a LinearRegression, no qual obteve um coeficiente de correlação $r = 0,59$ utilizando-se 22 atributos de 1.868. A Tabela 20 apresenta os resultados obtidos pelos algoritmos que foram aplicados com atributo meta N.

Tabela 20: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de N com diferentes quantidades de atributos.

	Quantidades de Atributos					
	10		22		1868	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,50	87,47 %	0,52	86,50 %	0,58	83,05 %
M5Rules	0,34	94,20 %	0,34	95,11 %	0,40	93,61 %
LinearRegression	0,57	81,92 %	0,59	81,61 %	0,41	89,84 %

5.1.2.2 P - Fósforo

Para o P apenas um dos algoritmos obteve resultado acima do limiar de 0,70, sendo a SMOReg que obteve um coeficiente de correlação $r = 0,76$ utilizando todos os atributos. A Tabela 21 apresenta os resultados obtidos pelos algoritmos de regressão.

Tabela 21: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.

	Quantidades de Atributos					
	3		7		1868	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,36	107,16 %	0,63	84,34 %	0,76	65,97 %
M5Rules	0,27	95,88 %	0,57	81,89 %	0,54	85,28 %
LinearRegression	0,36	92,86 %	0,39	92,22 %	0,47	88,32 %

5.1.2.3 K - Potássio

Sendo o K atributo meta e dentre os algoritmos aplicados apenas a SMOReg atingiu correlação acima do limiar, obtendo um $r = 0,76$ utilizando os 1.868 atributos, caso este similar ao que ocorreu quando o atributo meta foi o P. A Tabela 22 apresenta os resultados obtidos.

Tabela 22: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.

	Quantidades de Atributos					
	3		8		1868	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,32	102,32 %	0,53	86,22 %	0,76	64,54 %
M5Rules	0,26	96,64 %	0,52	85,14 %	0,50	90,92 %
LinearRegression	0,40	91,55 %	0,46	88,34 %	0,40	95,45 %

5.1.2.4 Ca - Cálcio

Os algoritmos executados para Ca como atributo meta não obtiveram resultados acima do limiar de 0,70, sendo que o algoritmo que aproximou-se do valor foi a SMOReg que obteve um $r = 0,41$ utilizando-se do conjunto completo de atributos. A Tabela 23 apresenta os resultados obtidos pelos algoritmos.

Tabela 23: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.

Quantidades de Atributos						
Algoritmo	5		8		1868	
	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,08	99,28 %	0,21	97,38 %	0,41	90,90 %
M5Rules	0,07	99,57 %	0,05	99,75 %	0,26	101,60 %
Linear Regression	0,007	101,58 %	0,40	91,35 %	0,36	88,34 %

5.1.2.5 Mg - Magnésio

Para o Mg, nenhum dos algoritmos utilizados obtiveram resultados acima do limiar de 0,70, situação semelhante a que ocorreu quando o atributo meta foi o Ca. O algoritmo que teve um resultado que aproximou-se do limiar foi a SMOReg que obteve coeficiente de correlação de 0,65 utilizando-se de todos os atributos. A Tabela 24 apresenta os resultados obtidos para os três algoritmos utilizados.

Tabela 24: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.

Quantidades de Atributos						
Algoritmo	9		27		1868	
	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,46	90,82 %	0,49	90,07 %	0,65	75,90 %
M5Rules	0,42	92,83 %	0,42	91,20 %	0,47	88,04 %
LinearRegression	0,53	84,64 %	0,54	85,35 %	0,50	90,05 %

5.1.2.6 S - Enxofre

Para o S, apenas a SMOReg obteve um coeficiente de correlação acima do limiar de $r = 0,74$ utilizando todos os atributos, situação esta já apresentada por outros atributos meta anteriormente, no qual a SMOReg vem sendo o algoritmo a obter um valor para coeficiente de correlação acima de 0,70. Na Tabela 25 é apresentado os resultados obtidos pelos algoritmos.

Tabela 25: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.

	Quantidades de Atributos					
	4		7		1868	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,32	108,70 %	0,31	108,82 %	0,74	68,71 %
M5Rules	0,23	96,90 %	0,23	97,87 %	0,37	99,38 %
LinearRegression	0,29	95,35 %	0,31	95,03 %	0,30	96,46 %

5.1.2.7 Cu - Cobre

Considerando o Cu, dentre os algoritmos a SMOReg foi a única a obter coeficiente de correlação $r = 0,85$ acima do limiar, utilizando todos os atributos. Na Tabela 26 é apresentado os resultados obtidos pelos algoritmos de regressão.

Tabela 26: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.

	Quantidades de Atributos					
	3		6		1868	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,24	102 %	0,37	95,50 %	0,85	53,07 %
M5Rules	0,29	95,51 %	0,33	93,92 %	0,64	76,58 %
LinearRegression	0,46	88,18 %	0,46	88,15 %	0,38	92,12 %

5.1.2.8 Mn - Manganês

Para o Mn, o algoritmo SMOReg obteve coeficiente de correlação acima do limiar, sendo $r = 0,77$ usando todos os atributos. A Tabela 27 apresenta os resultados obtidos após aplicação dos algoritmos.

Tabela 27: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.

	Quantidades de Atributos					
	2		4		1868	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,47	91,86 %	0,50	89 %	0,77	63,42 %
M5Rules	0,29	95,20 %	0,27	95,99 %	0,52	86,61 %
LinearRegression	0,46	88,10 %	0,53	84,21 %	0,49	82,52 %

5.1.2.9 Fe - Ferro

Considerando o Fe, os algoritmos não atingiram o limiar para coeficientes de correlação, exceto a SMOReg que utilizando-se de todos os atributos obteve um coeficiente de correlação $r = 0,81$. Na Tabela 28 é possível visualizar os resultados obtidos pelos algoritmos.

Tabela 28: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Fe com diferentes quantidades de atributos.

		Quantidades de Atributos					
		1		3		1868	
Algoritmo	r	RRSE	r	RRSE	r	RRSE	
SMOReg	0,15	112,76 %	0,43	91,37 %	0,81	58,45 %	
M5Rules	0,005	100,79 %	0,31	95,25 %	0,46	89,25 %	
LinearRegression	0,03	99,90 %	0,42	90,21 %	0,39	92,12 %	

5.1.2.10 Zn - Zinco

Assumindo como meta o Zn, dois algoritmos obtiveram valores para coeficiente de correlação acima do limiar, a SMOReg que utilizando-se de todos os atributos obteve um $r = 0,84$, e a LinearRegression que utilizando-se de 23 atributos obteve $r = 0,74$. A Tabela 29 apresenta os resultados obtidos pelos algoritmos executados.

Tabela 29: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.

		Quantidades de Atributos					
		6		23		1868	
Algoritmo	r	RRSE	r	RRSE	r	RRSE	
SMOReg	0,50	90,16 %	0,50	89,06 %	0,84	53,97 %	
M5Rules	0,30	94,83 %	0,37	92,72 %	0,60	79,46 %	
LinearRegression	0,51	85,34 %	0,74	67,31 %	0,50	89,39 %	

O Quadro 8 apresenta os 23 atributos que o algoritmo LinearRegression utilizou para obter o coeficiente de correlação $r = 0,74$.

Quadro 8: Descrição dos 23 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,75 no intervalo de 407 a 3.584 cm^{-1} .

Comprimentos de Onda
407 - 417 - 426 - 434 - 444 - 451 - 457 - 463
469 - 476 - 484 - 494 - 500 - 2369 - 3233 - 3252
3292 - 3327 - 3362 - 3416 - 3557 - 3570 - 3584

5.1.2.11 B - Boro

Considerando B, nenhum dos algoritmos de regressão atingiu o limiar de correlação de 0,70. O algoritmo que mais aproximou-se do limiar foi a SMOREg que utilizando-se dos 1.868 atributos obteve um coeficiente de correlação de $r = 0,67$. A Tabela 30 apresenta os resultados obtidos pelos algoritmos.

Tabela 30: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.

	Quantidades de Atributos					
	4		37		1868	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOREg	0,25	98,84 %	0,30	99,53 %	0,67	74,34 %
M5Rules	0,15	98,88 %	0,25	97,03 %	0,41	92,81 %
LinearRegression	0,11	99,32 %	0,39	110,42 %	0,42	112,79 %

5.2 Análise de Grãos

Nesta seção são apresentados os resultados referentes as aplicações de algoritmos para geração de modelos de regressão, tais como a SMOREg, M5Rules e a LinearRegression, que foram aplicadas em base de dados gerada em espectrofotômetro NIR. A base de dados analisada nesta etapa é formada pelos valores obtidos através do laboratório de solos da UEPG, como também dos espectros das 285 amostras de grãos de soja lidas no equipamento NIR modelo DA 1650. As Tabelas 31, 32 e 33 apresentam uma síntese dos resultados utilizados, respectivamente, 10%, 20% e todos os atributos.

Tabela 31: Resultados da análise de grãos utilizando redução da dimensionalidade em 10 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOREg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	3	0,58	81,65	0,58	81,01	0,62	78,10
P	6	0,11	108,06	0,24	97,07	0,45	78,10
K	11	0,35	96,93	0,25	97,52	0,70	71,11
Ca	7	0,64	77,32	0,62	77,98	0,75	65,72
Mg	10	0,43	97,16	0,31	94,93	0,50	86,48
S	6	0,36	97,85	0,25	97,67	0,31	95,07
Cu	4	0,26	98,21	0,12	99,23	0,14	99,22
Mn	10	0,06	105,80	0,13	116,23	0,48	87,81
Fe	8	0,30	97,59	0,09	101,58	0,26	97,39
Zn	6	0,07	101,31	0,17	98,28	0,18	98,46
B	9	0,43	91,23	0,36	93,18	0,47	87,93

Tabela 32: Resultados da análise de grãos utilizando redução da dimensionalidade em 20 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	15	0,62	78,67	0,56	83,04	0,70	71,38
P	27	0,33	99,12	0,36	93,30	0,85	51,86
K	15	0,37	95,64	0,25	97,13	0,75	66,48
Ca	10	0,65	76,72	0,61	79,28	0,76	65,17
Mg	14	0,44	96,76	0,32	94,51	0,52	85,64
S	24	0,34	98,53	0,41	91,10	0,84	53,71
Cu	14	0,30	97,38	0,14	98,79	0,36	93,35
Mn	21	0,14	105,03	0,28	100,58	0,52	86,13
Fe	22	0,55	83,80	0,47	88,30	0,64	76,82
Zn	29	0,27	102,5	0,20	97,70	0,62	80,02
B	21	0,43	91,07	0,39	91,77	0,43	91,44

Tabela 33: Resultados da análise de grãos sem utilizar redução da dimensionalidade, totalizando 275 atributos.

Nutriente	Algoritmo					
	SMOReg		M5Rules		LinearRegression	
	r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	0,71	70,59	0,60	79,72	0,15	100
P	0,41	93,98	0,36	94,21	0,16	98,87
K	0,46	99,65	0,28	98,15	0,23	97,13
Ca	0,73	69,09	0,58	81,58	0,58	81,50
Mg	0,44	93,18	0,32	95,56	0,30	95,06
S	0,43	92,83	0,36	93,69	0,15	99,07
Cu	0,28	96,23	0,15	98,50	0,14	98,85
Mn	0,47	94,47	0,30	95,55	0,18	98,60
Fe	0,51	88,94	0,41	91,17	0,24	96,89
Zn	0,16	100,25	0,14	98,82	0,04	100,18
B	0,46	89,71	0,42	90,36	0,42	90,56

As subseções abaixo detalham os resultados apresentados.

5.2.1 N - Nitrogênio

Para o N, dois algoritmos obtiveram resultados para coeficientes de correlação acima do limiar de 0,70, o algoritmo SMOReg que obteve uma correlação $r = 0,71$ utilizando-se de todos atributos da base de grãos e o algoritmo LinearRegression que obteve um $r = 0,70$, mas diferentemente da SMOReg, utilizou 15 atributos. A Tabela 34 apresenta os resultados obtidos pelos algoritmos executados.

O algoritmo LinearRegression destacou-se pelo resultado obtido, no qual seu

Tabela 34: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de N com diferentes quantidades de atributos.

	Quantidades de Atributos					
	3		15		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,58	81,65 %	0,62	78,67 %	0,71	70,59 %
M5Rules	0,58	81,01 %	0,56	83,04 %	0,60	79,72 %
LinearRegression	0,62	78,10 %	0,70	71,38 %	0,15	100 %

coeficiente de correlação foi de $r = 0,70$ utilizando 15 atributos, resultado menor que a SMOReg que obteve $r = 0,71$ utilizando todos os atributos. No Quadro 9 é apresentado os atributos utilizados pela LinearRegression.

Quadro 9: Descrição dos 15 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,70 no intervalo de 1.102 a 1.338 nm.

Comprimentos de Onda
1.102 - 1.110 - 1.118 - 1.128 - 1.142 - 1.154 - 1.186 - 1.204 1.210 - 1.218 - 1.228 - 1.246 - 1.280 - 1.332 - 1.338

5.2.2 P - Fósforo

Para o nutriente P, apenas a LinearRegression obteve um resultado acima do limiar, atingindo um coeficiente de correlação $r = 0,85$ utilizando 27 atributos. Na Tabela 35 é apresentado os resultados obtidos pelos algoritmos.

Tabela 35: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.

	Quantidades de Atributos					
	6		27		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,11	108,06 %	0,33	99,12 %	0,41	93,98 %
M5Rules	0,24	97,07 %	0,36	93,30 %	0,36	94,21 %
LinearRegression	0,45	78,10 %	0,85	51,86 %	0,16	98,87 %

O Quadro 10 apresenta os 27 atributos utilizados pelo algoritmo LinearRegression que obteve coeficiente de correlação $r = 0,85$.

Quadro 10: Descrição dos 27 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,85 no intervalo de 1.100 a 1.486 nm.

Comprimentos de Onda
1.000 - 1.110 - 1.118 - 1.126 - 1.134 - 1.146 - 1.154 - 1.166 - 1.190
1.196 - 1.202 - 1.208 - 1.214 - 1.220 - 1.230 - 1.248 - 1.276 - 1.294
1.306 - 1.342 - 1.378 - 1.404 - 1.432 - 1.442 - 1.450 - 1.456 - 1.486

5.2.3 K - Potássio

Para o K apenas um dos algoritmos obteve resultado acima do limiar de 0,70, no caso a LinearRegression que obteve um coeficiente de correlação $r = 0,75$ utilizando 15 atributos. O segundo resultado que atingiu o limiar também foi obtido pela LinearRegression, com um $r = 0,70$ utilizando 11 atributos. A Tabela 36 apresenta os resultados obtidos pelos algoritmos.

Tabela 36: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.

	Quantidades de Atributos					
	11		15		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,35	96,93 %	0,37	95,64 %	0,46	99,65 %
M5Rules	0,25	97,52 %	0,25	97,13 %	0,28	98,15 %
LinearRegression	0,70	71,11 %	0,75	66,48 %	0,23	97,13 %

No Quadro 11 apresenta os 15 atributos utilizados pela LinearRegression para obter um coeficiente de correlação $r = 0,75$.

Quadro 11: Descrição dos 15 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,75 no intervalo de 1.180 a 1.480 nm.

Comprimentos de Onda
1.180 - 1.186 - 1.192 - 1.200 - 1.214
1.220 - 1.430 - 1.436 - 1.442 - 1.450
1.456 - 1.462 - 1.468 - 1.474 - 1.480

5.2.4 Ca - Cálcio

Admitindo meta o Ca dois algoritmos atingiram o limiar de 0,70 para coeficientes de correlação, sendo eles a SMOReg e a LinearRegression. A SMOReg obteve uma correlação $r = 0,73$ utilizando-se de todos os atributos. No caso da LinearRegression obteve um coeficiente de correlação $r = 0,76$ utilizando 10 atributos e quando utilizou-se de 7 atributos obteve uma correlação $r = 0,75$. A Tabela 37 apresenta os resultados obtidos pelos algoritmos de regressão.

Tabela 37: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.

	Quantidades de Atributos					
	7		10		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,64	77,32 %	0,65	76,72 %	0,73	69,09 %
M5Rules	0,62	77,98 %	0,61	79,28 %	0,58	81,58 %
LinearRegression	0,75	65,72 %	0,76	65,17 %	0,58	81,50 %

No Quadro 12 é apresentado os atributos utilizados pelo algoritmo LinearRegression para obter correlações $r = 0,75$ e $r = 0,76$, utilizando 7 e 10 atributos respectivamente.

Quadro 12: Descrição dos 10 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,76 no intervalo de 1.182 a 1.478 nm.

Comprimentos de Onda
1.182 - 1.414 - 1.420 - 1.426 - 1.434
1.440 - 1.450 - 1.456 - 1.468 - 1.478

5.2.5 Mg - Magnésio

Considerando Mg como meta, os algoritmos não apresentaram correlação que atingisse o limiar de 0,70. A Tabela 38 apresenta os resultados das correlações obtidas pelos algoritmos.

Tabela 38: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.

	Quantidades de Atributos					
	10		14		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,43	97,16 %	0,44	96,76 %	0,44	93,18 %
M5Rules	0,31	94,93 %	0,32	94,51 %	0,32	95,56 %
LinearRegression	0,50	86,48 %	0,52	85,64 %	0,30	95,06 %

5.2.6 S - Enxofre

Considerando S apenas o algoritmo LinearRegression obteve um coeficiente de correlação acima do limiar, sendo $r = 0,84$, utilizando 24 atributos. A Tabela 39 apresenta os valores de correlação obtidos pelos algoritmos.

O Quadro 13 descreve quais os 24 atributos utilizados pela LinearRegression quando obtido o coeficiente de correlação $r = 0,84$.

Tabela 39: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.

	Quantidades de Atributos					
	6		24		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,36	97,85 %	0,34	98,53 %	0,43	92,83 %
M5Rules	0,25	97,67 %	0,41	91,10 %	0,36	93,69 %
LinearRegression	0,31	95,07 %	0,84	53,71 %	0,15	99,07 %

Quadro 13: Descrição dos 24 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,84 no intervalo de 1.100 a 1.498 nm.

Comprimentos de Onda
1.100 - 1.108 - 1.120 - 1.126 - 1.134 - 1.172
1.182 - 1.194 - 1.200 - 1.212 - 1.224 - 1.238
1.254 - 1.266 - 1.280 - 1.288 - 1.308 - 1.316
1.322 - 1.330 - 1.424 - 1.430 - 1.450 - 1.498

5.2.7 Cu - Cobre

Assumindo como meta o Cu, não foi possível obter coeficientes de correlação que atingissem o limiar. Na Tabela 40 é apresentado os resultados obtidos pelos algoritmos.

Tabela 40: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.

	Quantidades de Atributos					
	4		14		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,26	98,21 %	0,30	97,38 %	0,28	96,23 %
M5Rules	0,12	99,23 %	0,14	98,79 %	0,15	98,50 %
LinearRegression	0,14	99,22 %	0,36	93,35 %	0,14	98,85 %

5.2.8 Mn - Manganês

Para o Mn como meta, os algoritmos não obtiveram resultados que atingissem o limiar para coeficientes de correlação de 0,70. A Tabela 41 apresenta os resultados obtidos pelos algoritmos de regressão.

Tabela 41: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.

Quantidades de Atributos						
		10		21		275
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,06	105,8 %	0,14	105,03 %	0,47	94,47 %
M5Rules	0,13	116,23 %	0,28	100,58 %	0,30	95,55 %
LinearRegression	0,48	87,81 %	0,52	86,13 %	0,18	98,60 %

5.2.9 Fe - Ferro

Considerando Fe como meta os algoritmos utilizados não atingiram o limiar de 0,70 para coeficientes de correlação. O algoritmo LinearRegression aproximou-se desse resultado obtendo um $r = 0,64$ utilizando 22 atributos. Na Tabela 42 é apresentado os resultados obtidos pelos algoritmos.

Tabela 42: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Fe com diferentes quantidades de atributos.

Quantidades de Atributos						
		8		22		275
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,30	97,59 %	0,55	83,8 %	0,51	88,94 %
M5Rules	0,09	101,58 %	0,47	88,3 %	0,41	91,17 %
LinearRegression	0,26	97,39 %	0,64	76,82 %	0,24	96,89 %

5.2.10 Zn - Zinco

Assumindo Zn como meta, apenas a LinearRegression obteve um resultado para o coeficiente de correlação que aproximou-se do limiar de 0,70, sendo que obteve $r = 0,62$ utilizando 29 atributos. A Tabela 43 apresenta os resultados obtidos pelos algoritmos.

Tabela 43: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.

Quantidades de Atributos						
		6		29		275
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,07	101,31 %	0,27	102,5 %	0,16	100,25 %
M5Rules	0,17	98,28 %	0,20	97,7 %	0,14	98,82 %
LinearRegression	0,18	98,46 %	0,62	80,02 %	0,04	100,18 %

5.2.11 B - Boro

Para o B os algoritmos obtiveram resultados abaixo do limiar 0,70 para seus respectivos coeficientes de correlação. A Tabela 44 apresenta os resultados obtidos pelos algoritmos.

Tabela 44: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.

	Quantidades de Atributos					
	9		21		275	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,43	91,23 %	0,43	91,07 %	0,46	89,71 %
M5Rules	0,36	93,18 %	0,39	91,77 %	0,42	90,36 %
LinearRegression	0,47	87,93 %	0,43	91,44 %	0,42	90,56 %

5.3 Análise Foliar - Dados Transformados

Nesta seção serão apresentados os resultados obtidos através da execução de algoritmos de regressão a partir de bases de dados contendo amostras de folhas e grãos de soja, sendo que os valores de refletância obtidos das leituras dos espectros foram transformados pelas funções quadrática, inversa e neperiana com objetivo de diminuir a diferença entre os valores de seus espectros.

5.3.1 Espectroscopia NIR

A partir de uma base de dados com 200 amostras de folhas de soja analisadas em um equipamento NIR modelo DA 1650, gerou-se 275 atributos contendo os valores de refletância espectral que foram transformados utilizando-se das funções quadrática, inversa e neperiana, totalizando ao final uma base de dados contendo 825 atributos correspondentes aos espectros transformados. As Tabelas 45, 46 e 47 apresentam uma síntese dos resultados utilizados, respectivamente, 10%, 20% e todos os atributos. Logo abaixo as subseções detalham os resultados apresentados.

Tabela 45: Resultados da análise foliar em dados transformados com redução da dimensionalidade em 10 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	4	0,37	96,16	0,18	99,47	0,42	90,41
P	17	0,43	107,41	0,84	55,04	0,87	49,77
K	51	0,52	88,32	0,81	58,68	0,90	43,94
Ca	9	0,19	97,92	0,28	100,39	0,37	92,58
Mg	7	0,54	84,45	0,52	85,22	0,65	75,65
S	28	0,67	101,78	0,75	67,01	0,86	51,28
Cu	10	0,56	85,43	0,53	84,32	0,60	79,54
Mn	19	0,53	89,39	0,48	87,51	0,58	81,35
Fe	22	0,45	92,53	0,31	101,98	0,17	83,51
Zn	8	0,26	109,70	0,51	85,47	0,55	83,51
B	43	0,50	86,30	0,22	97,96	0,65	78,87

Tabela 46: Resultados da análise foliar em dados transformados com redução da dimensionalidade em 20 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	40	0,69	73,23	0,54	83,92	0,67	74,27
P	66	0,81	91,86	0,81	59,15	0,93	35,81
K	104	0,74	69,29	0,80	60,24	0,90	43,86
Ca	14	0,22	96,78	0,24	96,55	0,41	91,46
Mg	12	0,61	79,15	0,59	80,12	0,66	75,28
S	71	0,76	87,44	0,77	63,58	0,87	50,51
Cu	21	0,59	82,43	0,58	80,65	0,60	80,03
Mn	27	0,54	86,81	0,49	87,08	0,70	71,80
Fe	63	0,59	82,94	0,48	87,87	0,24	96,33
Zn	11	0,26	109,55	0,59	80,31	0,54	84,33
B	68	0,49	87,10	0,40	91,13	0,56	88,82

Tabela 47: Resultados da análise foliar em dados transformados sem utilizar redução da dimensionalidade, totalizando 825 atributos.

Nutriente	Algoritmo					
	SMOReg		M5Rules		LinearRegression	
	r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	0,70	71,18	0,48	87,31	0,49	86,37
P	0,85	55,11	0,81	58,07	0,60	79,64
K	0,83	55,96	0,76	64,93	0,58	81,21
Ca	0,44	89,11	0,04	166,18	0,19	98,03
Mg	0,70	71,16	0,61	79,33	0,57	84,65
S	0,82	59,59	0,76	64,98	0,56	82,63
Cu	0,77	64	0,61	78,74	0,59	80,46
Mn	0,80	59,32	0,71	70,75	0,56	82,44
Fe	0,70	73,47	0,45	91,36	0,51	85,28
Zn	0,69	73,53	0,56	82,95	0,41	90,76
B	0,54	83,97	0,43	89,68	0,43	89,75

5.3.1.1 N - Nitrogênio

Considerando N como meta apenas um dos algoritmos obteve resultado que atingiu limiar de 0,70 para coeficientes de correlação, no caso a SMOReg, que obteve um $r = 0,70$ utilizando todos os atributos. A SMOReg obteve também um $r = 0,69$ utilizando 40 atributos. Outro algoritmo que aproximou-se do limiar utilizando 40 atributos foi a LinearRegression que obteve um $r = 0,67$. A Tabela 48 apresenta os resultados obtidos pelos algoritmos.

Tabela 48: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de N com diferentes quantidades de atributos.

Algoritmo	Quantidades de Atributos					
	4		40		825	
	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,37	96,16 %	0,69	73,23 %	0,70	71,18 %
M5Rules	0,18	99,47 %	0,54	83,92 %	0,48	87,31 %
LinearRegression	0,42	90,41 %	0,67	74,27 %	0,49	86,37 %

Em relação ao N, os resultados com os valores de refletância transformados foram semelhantes aqueles obtidos com dados não transformados. No caso da análise de dados transformados o algoritmo SMOReg utilizou três vezes mais atributos e obteve o mesmo $r = 0,70$ que obteve para dados não transformados. Em relação a LinearRegression, obteve um $r = 0,68$ utilizando-se de dados não transformados, uma diferença de 0,01 para o resultado com dados transformados ($r = 0,67$), neste caso utilizando 40 atributos ao invés de 18.

5.3.1.2 P - Fósforo

Considerando P os algoritmos obtiveram valores de correlação acima do limiar de 0,70. Utilizando todos os atributos a SMOReg e a M5Rules obtiveram coeficientes de correlação $r = 0,84$ e $r = 0,81$ respectivamente. Utilizando 66 atributos os algoritmos obtiveram correlações acima de 0,80 e quando 17 atributos foram utilizados os algoritmos M5Rules e LinearRegression obteram $r = 0,84$ e $r = 0,87$ respectivamente. Os resultados podem ser analisados através da Tabela 49.

Tabela 49: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.

	Quantidades de Atributos					
	17		66		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,43	107,41 %	0,81	91,86 %	0,85	55,11 %
M5Rules	0,84	55,04 %	0,81	59,15 %	0,81	58,07 %
LinearRegression	0,87	49,77 %	0,93	35,81 %	0,60	79,64 %

Os algoritmos obtiveram resultados acima de 0,80 para correlação, sendo que a LinearRegression obteve um $r = 0,93$ utilizando-se de 66 atributos. A utilização de dados transformados resultou relativamente em coeficientes de correlação melhores, como é possível verificar no caso da LinearRegression que fazendo uso de dados não transformados obteve $r = 0,69$ utilizando 6 atributos e um $r = 0,87$ fazendo uso de 17 atributos da base de dados transformados. O Quadro 14 apresenta os 17 atributos utilizados pelos algoritmos M5Rules e LinearRegression para obter $r = 0,84$ e $r = 0,87$ respectivamente.

Quadro 14: Descrição dos 17 atributos utilizados na execução dos algoritmos LinearRegression e M5Rules gerando correlações de 0,87 e 0,84, respectivamente, no intervalo de comprimentos de onda de 1.258 a 1.534 nm.

Comprimentos de Onda
Inv1.258 - Inv1.412 - Inv1.474 - Inv1.516 - Inv1.534 - Quad1.268
Quad1.422 - Quad1.430 - Quad1.468 - Quad1.514 - Quad1.524
Nep1.396 - Nep1.426 - Nep1.436 - Nep1.458 - Nep1.478 - Nep1.498

5.3.1.3 K - Potássio

Assumindo K os algoritmos obtiveram resultados de coeficientes de correlação acima de 0,70, sendo que a LinearRegression obteve os maiores resultados para este caso. Utilizando 104 e 51 atributos da base de dados transformada a LinearRegression obteve um $r = 0,90$. Em comparação aos resultados obtidos utilizando dados não transformados a LinearRegression melhorou para o caso em que utilizou 51 atributos em comparação ao

uso de 16 da base não transformada e piorou quando utilizou 104 em comparação as 35 utilizadas da base não transformada. A Tabela 50 apresenta os resultados obtidos pelos três algoritmos.

Tabela 50: Resultados obtidos a partir do algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.

Algoritmo	Quantidades de Atributos					
	51		104		825	
	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,52	88,32 %	0,74	69,29 %	0,83	55,96 %
M5Rules	0,81	58,68 %	0,80	60,24 %	0,76	64,93 %
LinearRegression	0,90	43,94 %	0,90	43,86 %	0,58	81,21 %

A LinearRegression obteve o melhor resultado, $r = 0,90$ utilizando 51 atributos.

O Quadro 15 apresenta os atributos que foram utilizados.

Quadro 15: Descrição dos 19 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,90 no intervalo de comprimentos de onda de 1.102 a 1.536 nm.

Comprimentos de Onda
Inv1.244 - Inv1.262 - Inv1.268 - Inv1.282 - Inv1.324 - Inv1.420
Inv1.456 - Inv1.464 - Inv1.488 - Inv1.512 - Inv1.536 - Quad1.118
Quad1.248 - Quad1.258 - Quad1.274 - Quad1.290 - Quad1.302
Quad1.312 - Quad1.322 - Quad1.330 - Quad1.396 - Quad1.406
Quad1.412 - Quad1.426 - Quad1.430 - Quad1.436 - Quad1.448
Quad1.458 - Quad1.478 - Quad1.490 - Quad1.504 - Quad1.516
Quad1.522 - Quad1.528 - Nep1.102 - Nep1.108 - Nep1.122
Nep1.128 - Nep1.246 - Nep1.278 - Nep1.296 - Nep1.302
Nep1.326 - Nep1.392 - Nep1.400 - Nep1.414 - Nep1.420
Nep1.490 - Nep1.500 - Nep1.512 - Nep1.518

5.3.1.4 Ca - Cálcio

Considerando Ca os algoritmos não obtiveram valores para coeficientes de correlação que atingissem o limiar. A Tabela 51 apresenta os resultados obtidos.

Tabela 51: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.

Algoritmo	Quantidades de Atributos					
	9		14		825	
	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,19	97,92 %	0,22	96,78 %	0,44	89,11 %
M5Rules	0,28	100,39 %	0,24	96,55 %	0,04	166,18 %
LinearRegression	0,37	92,58 %	0,41	91,46 %	0,19	98,03 %

5.3.1.5 Mg - Magnésio

Assumindo Mg como meta, apenas a SMOReg atingiu o limiar estipulado e obteve um coeficiente de correlação $r = 0,70$ utilizando todos os atributos. A utilização de uma base transformada não gerou maiores resultados de correlação em comparação aqueles obtidos pela base não transformada. A Tabela 52 apresenta os resultados obtidos pelos algoritmos.

Tabela 52: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.

Algoritmo	Quantidades de Atributos					
	7		12		825	
	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,54	84,45 %	0,61	79,15 %	0,70	71,16 %
M5Rules	0,52	85,22 %	0,59	80,12 %	0,61	79,33 %
LinearRegression	0,65	75,65 %	0,66	75,28 %	0,57	84,65 %

5.3.1.6 S - Enxofre

Admitindo S, os algoritmos obtiveram resultados para coeficientes de correlação acima do limiar, sendo que o algoritmo LinearRegression obteve $r = 0,87$ utilizando 71 atributos. Fazendo uso de uma quantidade menor de atributos, 28, os algoritmos M5Rules e LinearRegression obtiveram valores $r = 0,75$ e $r = 0,86$ respectivamente. A Tabela 53 apresenta os resultados obtidos pelos algoritmos.

Tabela 53: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.

Algoritmo	Quantidades de Atributos					
	28		71		825	
	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,67	101,78 %	0,76	87,44 %	0,82	59,59 %
M5Rules	0,75	67,01 %	0,77	63,58 %	0,76	64,98 %
LinearRegression	0,86	51,28 %	0,87	50,51 %	0,56	82,63 %

A LinearRegression obteve os maiores resultados para coeficientes de correlação, tanto utilizando 28 atributos como para 71, sendo $r = 0,86$ e $r = 0,87$ respectivamente. No Quadro 16 é apresentado os 28 atributos utilizados.

Quadro 16: Descrição dos 28 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,86 no intervalo de comprimentos de onda de 1.240 a 1.542 nm.

Comprimentos de Onda
Inv1.250 - Inv1.262 - Inv1.272 - Inv1.280 - Inv1.386 - Inv1.400 Inv1.492 - Inv1.510 - Inv1.530 - Quad1.256 - Quad1.276 Quad1.292 - Quad1.298 - Quad1.404 - Quad1.418 - Quad1.430 Quad1.458 - Quad1.516 - Quad1.522 - Quad1.542 - Nep1.240 Nep1.280 - Nep1.418 - Nep1.448 - Nep1.454 Nep1.482 - Nep1.490 - Nep1.524

5.3.1.7 Cu - Cobre

Considerando o Cu apenas um dentre os algoritmos atingiu correlação acima do limiar, sendo a SMOReg que obteve um coeficiente de correlação $r = 0,77$ utilizando todos os atributos. Em comparação ao resultado obtido pela SMOReg utilizando dados não transformados, o coeficiente de correlação teve um aumento de 0,07 mas, envolveram uma quantidade maior de atributos. A Tabela 54 apresenta os resultados obtidos pelos algoritmos.

Tabela 54: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.

	Quantidades de Atributos					
	10		21		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,56	85,43 %	0,59	82,43 %	0,77	64 %
M5Rules	0,53	84,32 %	0,58	80,65 %	0,61	78,74 %
LinearRegression	0,60	79,54 %	0,60	80,03 %	0,59	80,46 %

5.3.1.8 Mn - Manganês

Assumindo Mn, os algoritmos obtiveram resultados para seus respectivos coeficientes de correlação acima do limiar de 0,70, sendo que a SMOReg e a M5Rules obtiveram $r = 0,80$ e $r = 0,71$ respectivamente, utilizando todos os atributos. A SMOReg obteve um valor maior de correlação utilizando base transformada em comparação ao seu resultado utilizando a base não transformada, sendo que a M5Rules obteve um menor valor fazendo uso de dados transformados. A Tabela 55 apresenta os resultados obtidos pelos algoritmos.

A LinearRegression obteve um $r = 0,70$ utilizando 27 atributos, sendo este um resultado maior em comparação ao uso de base não transformada, mas deve-se levar em consideração que o resultado menor, um $r = 0,68$ foi utilizando 7 atributos. No Quadro 21

Tabela 55: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.

	Quantidades de Atributos					
	19		27		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,53	89,39 %	0,54	86,81 %	0,80	59,32 %
M5Rules	0,48	87,51 %	0,49	87,08 %	0,71	70,75 %
LinearRegression	0,58	81,35 %	0,70	71,80 %	0,56	82,44 %

é apresentado os atributos utilizados pela LinearRegression.

Quadro 17: Descrição dos 27 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,70 no intervalo de comprimentos de onda de 1.100 a 1.158 nm.

Comprimentos de Onda
Inv1.100 - Inv1.108 - Inv1.114 - Inv1.120 - Inv1.126 - Inv1.132 Inv1.138 - Inv1.144 - Inv1.150 - Inv1.156 - Quad1.100 - Quad1.108 Quad1.114 - Quad1.120 - Quad1.126 - Quad1.132 - Quad1.140 Quad1.148 - Nep1.100 - Nep1.106 - Nep1.114 - Nep1.120 Nep1.126 - Nep1.134 - Nep1.140 - Nep1.148 - Nep1.158

5.3.1.9 Fe - Ferro

Considerando Fe, apenas o algoritmo SMOReg obteve coeficiente de correlação acima do limiar, $r = 0,70$ utilizando todos os atributos, resultado este maior do que aquele obtido fazendo uso de base não transformada ($r = 0,67$) utilizando 275 atributos. A Tabela 56 apresenta os resultados obtidos.

Tabela 56: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Fe com diferentes quantidades de atributos.

	Quantidades de Atributos					
	22		63		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,45	92,53 %	0,59	82,94 %	0,70	73,47 %
M5Rules	0,31	101,98 %	0,48	87,87 %	0,45	91,36 %
LinearRegression	0,17	83,51 %	0,24	96,33 %	0,51	85,28 %

5.3.1.10 Zn - Zinco

Assumindo Zn, nenhum algoritmo atingiu o limiar esperado de 0,70, sendo que a SMOReg obteve um resultado de $r = 0,69$ que aproximou-se do limiar utilizando todos os atributos. Este resultado foi melhor do que aquele obtido utilizando base não transformada

mesmo não atingindo o limiar. Na Tabela 57 é apresentado os resultados obtidos pelos algoritmos.

Tabela 57: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.

	Quantidades de Atributos					
	8		11		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,26	109,70 %	0,26	109,55 %	0,69	73,53 %
M5Rules	0,51	85,47 %	0,59	80,31 %	0,56	82,95 %
LinearRegression	0,55	83,51 %	0,54	84,33 %	0,41	90,76 %

5.3.1.11 B - Boro

Considerando B os algoritmos não obtiveram resultados que atingissem o limiar esperado de 0,70, sendo que o resultado que aproximou-se foi obtido pela LinearRegression, $r = 0,65$ utilizando 43 atributos. Os resultados obtidos mostraram-se maiores do que os obtidos pelo uso de base não transformada mas não atingiram o limiar. A Tabela 58 apresenta os resultados obtidos.

Tabela 58: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.

	Quantidades de Atributos					
	43		68		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,50	86,30 %	0,49	87,10 %	0,54	83,97 %
M5Rules	0,22	97,96 %	0,40	91,13 %	0,43	89,68 %
LinearRegression	0,65	78,87 %	0,56	88,82 %	0,43	89,75 %

5.3.2 Espectroscopia MIR

Os resultados apresentados nesta seção foram obtidos a partir de aplicações de três algoritmos em bases de dados espectrais de 231 amostras de folhas de soja analisadas em um espectrofotômetro F-TIR modelo IRPrestige-21 entre os números de onda do 4.000 cm^{-1} até 400 cm^{-1} .

Inicialmente a base possuía 1868 atributos correspondentes aos números de onda do infravermelho médio. Para este caso foram efetuados transformações em tais valores aplicando as funções quadrática, inversa e neperiana, sendo que a quantidade total de atributos após as transformações passou para 5.604. As Tabelas 59, 60 e 61 apresentam

uma síntese dos resultados utilizados, respectivamente, 10%, 20% e todos os atributos. Logo abaixo as próximas subseções detalham os resultados apresentados.

Tabela 59: Resultados da análise foliar em dados transformados com redução da dimensionalidade em 10 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	8	0,42	92,38	0,31	95,74	0,59	80,95
P	3	0,39	106,53	0,28	95,69	0,40	91,60
K	4	0,34	101,75	0,28	95,72	0,41	90,93
Ca	6	0,09	99,04	0,07	99,52	0,07	99,79
Mg	12	0,34	94,14	0,40	91,97	0,54	84,05
S	4	0,30	108,81	0,31	94,60	0,31	94,65
Cu	4	0,11	103,64	0,38	91,74	0,44	89,20
Mn	3	0,40	94,39	0,33	93,80	0,52	85,07
Fe	6	0,15	101,51	0,18	98,48	0,06	105,12
Zn	14	0,39	97,14	0,43	89,60	0,52	87,51
B	2	0,02	101,66	0,16	100,51	0,12	99,39

Tabela 60: Resultados da análise foliar em dados transformados com redução da dimensionalidade em 20 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	50	0,54	85,31	0,43	92,16	0,52	88,55
P	6	0,39	106,35	0,23	97,08	0,37	93,09
K	11	0,42	96,69	0,45	91,12	0,51	85,96
Ca	22	0,34	93,49	0,37	92,85	0,36	92,68
Mg	29	0,38	92,72	0,36	93,34	0,45	96,93
S	8	0,11	108,34	0,29	95,44	0,25	97,96
Cu	8	0,16	102,73	0,32	95,36	0,55	82,88
Mn	9	0,51	89,48	0,33	94,83	0,51	86,34
Fe	17	0,03	108,32	0,39	94,76	0,57	86,67
Zn	53	0,73	69,52	0,60	79,06	0,81	60,62
B	10	0,27	95,76	0,10	99,47	0,49	87,27

Tabela 61: Resultados da análise foliar em dados transformados sem utilizar redução da dimensionalidade, totalizando 5604 atributos.

Nutriente	Algoritmo					
	SMOReg		M5Rules		LinearRegression	
	r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	0,58	83,40	0,42	92,90	0,45	93,75
P	0,79	62,39	0,61	78,84	0,32	95,34
K	0,77	63,92	0,62	78,88	0,49	83,50
Ca	0,40	95,27	0,33	95,26	0,40	95,35
Mg	0,62	80,20	0,46	89,01	0,49	93,87
S	0,77	64,53	0,48	90,60	0,29	96,93
Cu	0,87	49,98	0,55	84,33	0,41	95,45
Mn	0,79	61,59	0,55	82,75	0,48	88,12
Fe	0,83	55,32	0,42	91,06	0,23	90,32
Zn	0,85	51,95	0,57	82,34	0,49	90,34
B	0,66	77,37	0,48	87,16	0,23	92,71

5.3.2.1 N - Nitrogênio

Considerando o N como meta, os algoritmos não obtiveram resultados que atin- gissem o limiar estipulado de 0,70 para coeficientes de correlação, sendo que o maior re- sultado encontrado foi pelo algoritmo LinearRegression que utilizando 8 atributos obteve um $r = 0,59$, sendo este também o maior resultado obtido por bases não transformadas. Na Tabela 62 é apresentado os resultados obtidos pelos algoritmos.

Tabela 62: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de N com diferentes quantidades de atributos.

Algoritmo	Quantidades de Atributos					
	8		50		5604	
	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,42	92,38 %	0,54	85,31 %	0,58	83,40 %
M5Rules	0,31	95,74 %	0,43	92,16 %	0,42	92,90 %
LinearRegression	0,59	80,95 %	0,52	88,55 %	0,45	93,75 %

5.3.2.2 P - Fósforo

Assumindo meta o P, o limiar para coeficientes de correlação estipulado foi atin- gido pelo algoritmo SMOReg quando este utilizou-se de todos os atributos, obtendo um $r = 0,79$, resultado este maior que o encontrado utilizando bases não transformadas, mas neste caso o resultado menor encontrado utilizou uma quantidade menor de atributos, 1.868 e obteve $r = 0,76$. A Tabela 63 apresenta os resultados obtidos.

Tabela 63: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.

Quantidades de Atributos						
		3		6		5604
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,39	106,53 %	0,39	106,35 %	0,79	62,39 %
M5Rules	0,28	95,69 %	0,23	97,08 %	0,61	78,84 %
LinearRegression	0,40	91,60 %	0,37	93,09 %	0,32	95,34 %

5.3.2.3 K - Potássio

Considerando K, apenas um algoritmo obteve resultado acima do limiar estipulado, a SMOReg com $r = 0,77$ utilizando todos os atributos, resultado este 0,01 acima daquele obtido pelo mesmo algoritmo utilizando base não transformada, mas no caso fazendo uso de 1.868 atributos. A Tabela 64 apresenta os resultados obtidos pelos algoritmos.

Tabela 64: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.

Quantidades de Atributos						
		4		11		5604
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,34	101,75 %	0,42	96,69 %	0,77	63,92 %
M5Rules	0,28	95,72 %	0,45	91,12 %	0,62	78,88 %
LinearRegression	0,41	90,93 %	0,51	85,96 %	0,49	83,50 %

5.3.2.4 Ca - Cálcio

Para Ca como meta os algoritmos não atingiram o limiar, obtendo resultados inferiores aqueles obtidos utilizando bases não transformadas. A Tabela 65 apresenta os resultados obtidos.

Tabela 65: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.

Quantidades de Atributos						
		6		22		5604
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,09	99,04 %	0,34	93,49 %	0,40	95,27 %
M5Rules	0,07	99,52 %	0,37	92,85 %	0,33	95,26 %
LinearRegression	0,07	99,79 %	0,36	92,68 %	0,40	95,35 %

5.3.2.5 Mg - Magnésio

Considerando Mg, os algoritmos não obtiveram valores para correlação que atin- gissem o limiar, sendo que o algoritmo SMOReg aproximou-se deste valor obtendo um $r = 0,62$ utilizando todos os atributos, resultado este menor do que aquele obtido utilizando bases não transformadas. A Tabela 66 apresenta os resultados obtidos.

Tabela 66: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.

	Quantidades de Atributos					
	12		29		5604	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,34	94,14 %	0,38	92,72 %	0,62	80,20 %
M5Rules	0,40	91,97 %	0,36	93,34 %	0,46	89,01 %
LinearRegression	0,54	84,05 %	0,45	96,93 %	0,49	93,87 %

5.3.2.6 S - Enxofre

Assumindo S como meta apenas a SMOReg obteve um valor para coeficientes de correlação acima do limiar, no qual obteve $r = 0,77$ utilizando todos os atributos. Re- sultado este 0,03 acima do obtido através de bases não transformadas. Na Tabela 67 é apresentado os resultados obtidos pelos algoritmos.

Tabela 67: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.

	Quantidades de Atributos					
	4		8		5604	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,30	108,81 %	0,11	108,34 %	0,77	64,53 %
M5Rules	0,31	94,60 %	0,29	95,44 %	0,48	90,60 %
LinearRegression	0,31	94,65 %	0,25	97,96 %	0,29	96,93 %

5.3.2.7 Cu - Cobre

Considerando Cu, apenas um algoritmo atingiu limiar estipulado, no caso a SMO- Reg que obteve coeficiente de correlação $r = 0,87$ utilizando todos os atributos, resultado este 0,02 maior que aquele encontrado em bases não transformadas. Na Tabela 68 é apre- sentado os resultados obtidos pelos algoritmos.

Tabela 68: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.

Quantidades de Atributos						
		4		8		5604
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,11	103,64 %	0,16	102,73 %	0,87	49,98 %
M5Rules	0,38	91,74 %	0,32	95,36 %	0,55	84,33 %
LinearRegression	0,44	89,20 %	0,55	82,88 %	0,41	95,45 %

5.3.2.8 Mn - Manganês

Para Mn o algoritmo SMOReg obteve o único resultado para coeficientes de correlação superior ao limiar, no caso $r = 0,79$ fazendo uso de todos os atributos. Resultado este 0,02 maior que aquele obtido por base não transformada. Na Tabela 69 é apresentado os resultados obtidos pelos algoritmos.

Tabela 69: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.

Quantidades de Atributos						
		3		9		5604
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,40	94,39 %	0,51	89,48 %	0,79	61,59 %
M5Rules	0,33	93,80 %	0,33	94,83 %	0,55	82,75 %
LinearRegression	0,52	85,07 %	0,51	86,34 %	0,48	88,12 %

5.3.2.9 Fe - Ferro

Considerando Fe apenas o algoritmo SMOReg obteve valores para correlação acima do limiar, no qual obteve $r = 0,83$ utilizando todos os atributos. Este resultado foi 0,02 maior que aquele obtido pela SMOReg utilizando todos os atributos da base não transformada. A Tabela 70 apresenta os resultado obtidos pelos algoritmos.

Tabela 70: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativas do teor de Fe com diferentes quantidades de atributos.

Quantidades de Atributos						
		6		17		5604
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,15	101,51 %	0,03	108,32 %	0,83	55,32 %
M5Rules	0,18	98,48 %	0,39	94,76 %	0,42	91,06 %
LinearRegression	0,06	105,12 %	0,57	86,67 %	0,23	90,32 %

5.3.2.10 Zn - Zinco

Assumindo Zn, apenas o algoritmo M5Rules não atingiu o limiar estipulado, sendo que a SMOReg obteve coeficientes de correlação $r = 0,73$ e $r = 0,85$ utilizando 53 e 5.604 atributos, respectivamente. A LinearRegression também obteve valores para correlação acima do limiar, sendo $r = 0,81$ fazendo uso de 53 atributos. Em relação ao uso de bases transformadas, os resultados foram maiores para SMOReg com diferença de 0,23 e para LinearRegression teve um aumento de 0,06. A Tabela 71 apresenta os resultados obtidos pelos algoritmos.

Tabela 71: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.

	Quantidades de Atributos					
	14		53		5604	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,39	97,14 %	0,73	69,52 %	0,85	51,95 %
M5Rules	0,43	89,60 %	0,60	79,06 %	0,57	82,34 %
LinearRegression	0,52	87,51 %	0,81	60,62 %	0,49	90,34 %

Dois algoritmos obtiveram valores para coeficientes de correlação acima de 0,80, SMOReg e LinearRegression, o primeiro utilizando-se de todos os atributos obteve $r = 0,85$ e o segundo utilizando 53 obteve $r = 0,81$. O Quadro 18 apresenta os 53 atributos utilizados pela LinearRegression.

Quadro 18: Descrição dos 53 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,81 no intervalo de números de onda de 399 a 3628cm⁻¹.

Comprimentos de Onda
Inv399 - Inv405 - Inv413 - Inv420 - Inv426 - Inv438 - Inv446
Inv451 - Inv461 - Inv467 - Inv478 - Inv484 - Inv490 - Inv496
Inv503 - Qua449 - Qua650 - Qua1063 - Qua1231 - Qua1254
Qua1416 - Qua1508 - Qua1620 - Qua1628 - Qua1634 - Qua1655
Qua1672 - Qua1821 - Qua1888 - Qua1911 - Qua1950 - Qua1971
Qua2006 - Qua2145 - Qua2160 - Qua2313 - Qua2347
Qua2918 - Qua2924 - Qua3034 - Qua3167 - Qua3273
Qua3319 - Qua3348 - Qua3395 - Qua3418 - Qua3464
Qua3530 - Qua3543 - Qua3628 - Nep447 - Nep463

5.3.2.11 B - Boro

Considerando B os algoritmos não obtiveram resultados para seus respectivos coeficientes de correlação que atingissem o limiar de 0,70. A SMOReg foi o algoritmo que

aproximou-se do limiar obtendo um $r = 0,66$ utilizando todos os atributos. Este resultado foi 0,01 menor que aquele obtido quando utilizado bases não transformadas. A Tabela 72 apresenta os resultados obtidos pelos algoritmos.

Tabela 72: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.

	Quantidades de Atributos					
	2		10		5604	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOREg	0,02	101,66 %	0,27	95,76 %	0,66	77,37 %
M5Rules	0,16	100,51 %	0,10	99,47 %	0,48	87,16 %
LinearRegression	0,12	99,39 %	0,49	87,27 %	0,23	92,71 %

5.4 Análise de Grãos - Dados Transformados

Nesta seção são apresentados os resultados referentes a aplicação de algoritmos para geração de modelos de regressão, tais como a SMOREg, M5Rules e a LinearRegression que foram aplicadas em base de dados geradas em espectrofotômetro NIR modelo 1650 que possui comprimentos de onda de 1.100 nm até 1.650 nm.

A base de dados analisada nesta etapa é formada pelos valores obtidos através do laboratório de solos da UEPG, como também dos espectros das 285 amostras de grãos de soja lidas no equipamento NIR. Inicialmente a base continha 275 atributos e após efetuada transformações nestes valores com objetivo de diminuir a diferença entre os mesmos, foi utilizada as funções inversa, quadrática e neperiana, sendo que no fim os atributos totalizaram 825.

As Tabelas 73, 74 e 75 apresentam uma síntese dos resultados utilizados, respectivamente, 10%, 20% e todos os atributos. Logo abaixo as subseções detalham os resultados apresentados.

Tabela 73: Resultados da análise de grãos em dados transformados com redução da dimensionalidade em 10 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	12	0,57	82,11	0,57	82,49	0,70	71,72
P	16	0,04	108,50	0,36	92,97	0,71	71,01
K	14	0,22	105,52	0,23	98	0,69	72,54
Ca	10	0,66	75,56	0,65	76,02	0,73	68,13
Mg	22	0,44	95,13	0,36	93,88	0,57	82,75
S	21	0,37	107,34	0,28	98,66	0,78	62,21
Cu	12	0,13	100,76	0,10	100,24	0,34	95,55
Mn	7	0,05	106,10	0,16	98,57	0,15	98,61
Fe	9	0,01	103,38	0,11	101,17	0,21	97,68
Zn	4	0,10	100,11	0,19	97,99	0,20	96,67
B	13	0,45	89,89	0,41	91,41	0,42	92,38

Tabela 74: Resultados da análise de grãos em dados transformados com redução da dimensionalidade em 20 % dos atributos.

Nutriente	Atributos	Algoritmo					
		SMOReg		M5Rules		LinearRegression	
		r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	41	0,63	77,22	0,63	77,89	0,65	80,50
P	37	0,07	108,54	0,35	93,87	0,85	53,64
K	24	0,22	105,22	0,32	94,73	0,77	64,17
Ca	24	0,70	72,64	0,65	75,87	0,77	63,46
Mg	47	0,39	94,57	0,28	97,81	0,57	84,16
S	46	0,36	107,53	0,29	96,78	0,83	55,89
Cu	58	0,22	97,57	0,12	99,28	0,57	89,40
Mn	15	0,08	106,40	0,08	102,27	0,28	99,07
Fe	12	0,04	107,59	0,06	104,55	0,14	100,48
Zn	19	0,08	101,77	0,20	97,66	0,40	106,60
B	24	0,46	89,76	0,37	92,98	0,43	92,06

Tabela 75: Resultados da análise de grãos sem utilizar redução da dimensionalidade, totalizando 825 atributos.

Nutriente	Algoritmo					
	SMOReg		M5Rules		LinearRegression	
	r	RRSE (%)	r	RRSE (%)	r	RRSE (%)
N	0,72	69,14	0,61	79,14	0,59	86,89
P	0,60	102,07	0,36	93,95	0,31	94,92
K	0,59	93,80	0,29	92,26	0,21	98,02
Ca	0,74	67,74	0,57	82,60	0,58	81,46
Mg	0,49	91,57	0,29	96,31	0,31	94,76
S	0,64	102,79	0,31	99,04	0,56	70,34
Cu	0,25	98,08	0,12	100,77	0,41	92,31
Mn	0,50	92,52	0,30	95,87	0,10	100,09
Fe	0,42	92,93	0,37	92,42	0,11	105,09
Zn	0,24	98,48	0,20	98,15	0,25	99,97
B	0,50	87,69	0,37	93,29	0,39	95,56

5.4.1 N - Nitrogênio

Considerando N, dois algoritmos atingiram o limiar de 0,70, a SMOReg que obteve coeficiente de correlação $r = 0,72$ utilizando todos os atributos e a LinearRegression que obteve $r = 0,70$ utilizando 12 atributos. A Tabela 76 apresenta os resultados obtidos.

Tabela 76: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativas do teor de N com diferentes quantidades de atributos.

Algoritmo	Quantidades de Atributos					
	12		41		825	
	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,57	82,11 %	0,63	77,22 %	0,72	69,14 %
M5Rules	0,57	82,49 %	0,63	77,89 %	0,61	79,14 %
LinearRegression	0,70	71,72 %	0,65	80,50 %	0,59	86,89 %

A partir dos resultados obtidos através da base de dados transformada, apenas para a SMOReg houve mudança nos resultados obtidos, uma diferença de 0,01. No Quadro 19 é apresentado 12 atributos que a LinearRegression utilizou para obter $r = 0,70$.

Quadro 19: Descrição dos 12 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,70 no intervalo de comprimentos de onda de 1.102 a 1.322 nm.

Comprimentos de Onda
Inv1.240 - Inv1.266 - Inv1.288 - Inv1.300
Inv1.320 - Qua1.102 - Qua1.114 - Nep1.110
Nep1.134 - Nep1.144 - Nep1.308 - Nep1.322

5.4.2 P - Fósforo

Assumindo P como meta, apenas a LinearRegression obteve resultados acima do limiar para coeficientes de correlação, sendo $r = 0,71$ utilizando 16 atributos e $r = 0,85$ utilizando 37 atributos. A Tabela 77 apresenta os resultados obtidos pelos algoritmos.

Tabela 77: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de P com diferentes quantidades de atributos.

	Quantidades de Atributos					
	16		37		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,04	108,50 %	0,07	108,54 %	0,60	102,07 %
M5Rules	0,36	92,97 %	0,35	93,87 %	0,36	93,95 %
LinearRegression	0,71	71,01 %	0,85	53,64 %	0,31	94,92 %

No Quadro 20 é apresentado os 16 atributos utilizados pela LinearRegression.

Quadro 20: Descrição dos 16 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,71 no intervalo de comprimentos de onda de 1.104 a 1.478 nm.

Comprimentos de Onda
Inv1.188 - Inv1.200 - Inv1.218 - Qua1.114
Qua1.190 - Qua1.198 - Qua1.220 - Qua1.452
Qua1.468 - Nep1.104 - Nep1.188 - Nep1.200
Nep1.446 - Nep1.458 - Nep1.470 - Nep1.478

5.4.3 K - Potássio

Considerando K, apenas um algoritmo obteve resultado para coeficientes de correlação acima do limiar, sendo este a LinearRegression que obteve um $r = 0,77$ fazendo uso de 24 atributos. A LinearRegression obteve também um valor de $r = 0,69$ utilizando 14 atributos. A Tabela 78 apresenta os resultados obtidos.

Tabela 78: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de K com diferentes quantidades de atributos.

	Quantidades de Atributos					
	14		24		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,22	105,52 %	0,22	105,22 %	0,59	93,80 %
M5Rules	0,23	98 %	0,32	94,73 %	0,29	92,26 %
LinearRegression	0,69	72,54 %	0,77	64,17 %	0,21	98,02 %

No Quadro 21 é apresentado os 24 atributos utilizados pela LinearRegression.

Quadro 21: Descrição dos 24 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,77 no intervalo de comprimentos de onda de 1.118 a 1.486 nm.

Comprimentos de Onda
Inv1.178 - Inv1.188 - Inv1.202 - Inv1.420 - Inv1.432 - Inv1.450
Inv1.458 - Inv1.466 - Inv1.478 - Qua1.118 - Qua1.196 - Qua1.204
Qua1.416 - Qua1.448 - Qua1.468 - Qua1.486 - Nep1.210 - Nep1.216
Nep1.228 - Nep1.432 - Nep1.438 - Nep1.444 - Nep1.464 - Nep1.486

5.4.4 Ca - Cálcio

Assumindo Ca, tanto o algoritmo SMOReg como a LinearRegression obtiveram resultados para seus respectivos coeficientes de correlação acima no limiar. A SMOReg obteve $r = 0,70$ e $r = 0,74$ utilizando-se de 24 atributos e 825 atributos, respectivamente. No caso da RegressãoLinear, esta obteve $r = 0,73$ e $r = 0,77$ utilizando 10 e 24 atributos, respectivamente. A Tabela 79 apresenta os resultados obtidos pelos algoritmos.

Tabela 79: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Ca com diferentes quantidades de atributos.

	Quantidades de Atributos					
	10		24		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,66	75,56 %	0,70	72,64 %	0,74	67,74 %
M5Rules	0,65	76,02 %	0,65	75,87 %	0,57	82,60 %
LinearRegression	0,73	68,13 %	0,77	63,46 %	0,58	81,46 %

O Quadro 22 apresenta os 10 atributos utilizados pela LinearRegression, no qual obteve $r = 0,73$.

Quadro 22: Descrição dos 10 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,73 no intervalo de comprimentos de onda de 1.206 a 1.454 nm.

Comprimentos de Onda
Inv1.418 - Inv1.424 - Inv1.432 - Inv1.450 - Nep1.206
Nep1.212 - Nep1.418 - Nep1.426 - Nep1.442 - Nep1.454

5.4.5 Mg - Magnésio

Considerando Mg, não houve resultados para coeficientes de correlação que atinxissem o limiar 0,70, sendo que o algoritmo LinearRegression obteve o maior resultado para $r = 0,57$. Não houve resultados próximos ao limiar utilizando bases de dados transformados. A Tabela 80 apresenta os resultados obtidos.

Tabela 80: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mg com diferentes quantidades de atributos.

	Quantidades de Atributos					
	22		47		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,44	95,13 %	0,39	94,57 %	0,49	91,57 %
M5Rules	0,36	93,88 %	0,28	97,81 %	0,29	96,31 %
LinearRegression	0,57	82,75 %	0,57	84,16 %	0,31	94,76 %

5.4.6 S - Enxofre

Considerando S como meta, dentre os três algoritmos apenas a LinearRegression atingiu o limiar de 0,70 e obteve $r = 0,78$ utilizando 21 atributos. Resultado este maior do que aquele encontrado utilizando base não transformada. A Tabela 81 apresenta os resultados obtidos pelos algoritmos.

Tabela 81: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de S com diferentes quantidades de atributos.

	Quantidades de Atributos					
	21		46		825	
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,37	107,34 %	0,36	107,53 %	0,64	102,79 %
M5Rules	0,28	98,66 %	0,29	96,78 %	0,31	99,04 %
LinearRegression	0,78	62,21 %	0,83	55,89 %	0,56	70,34 %

No Quadro 23 é apresentado os 21 atributos utilizados pelo algoritmo LinearRegression que obteve $r = 0,78$.

Quadro 23: Descrição dos 21 atributos utilizados na execução do algoritmo LinearRegression gerando correlação de 0,78 no intervalo de comprimentos de onda de 1.102 a 1.476 nm.

Comprimentos de Onda
Inv1.198 - Inv1.206 - Inv1.422 - Inv1.430 - Inv1.438 - Inv1.456
Inv1.476 - Qua1.102 - Qua1.112 - Qua1.194 - Qua1.204
Qua1.212 - Qua1.440 Qua1.446 - Qua1.460 - Qua1.468
Nep1.202 - Nep1.212 - Nep1.220 - Nep1.438 - Nep.466

5.4.7 Cu - Cobre

Assumindo Cu não houve algoritmos que obtiveram resultados que atingissem o limiar de 0,70 para coeficientes de correlação, sendo que a LinearRegression obteve o maior resultado sendo $r = 0,57$ utilizando 58 atributos da base. A Tabela 82 apresenta os resultados encontrados pelos algoritmos.

Tabela 82: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Cu com diferentes quantidades de atributos.

Quantidades de Atributos						
		12		58		825
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,13	100,76 %	0,22	97,57 %	0,25	98,08 %
M5Rules	0,10	100,24 %	0,12	99,28 %	0,12	100,77 %
LinearRegression	0,34	95,55 %	0,57	89,40 %	0,41	92,31 %

5.4.8 Mn - Manganês

Considerando Mn, dentre os algoritmos não houve resultados que atingissem o limiar de 0,70 para coeficientes de correlação, sendo que o algoritmo LinearRegression obteve o maior resultado, $r = 0,28$. A utilização de bases de dados transformada não gerou resultados que aproximassem do limiar. Na Tabela 83 é apresentados os resultados obtidos.

Tabela 83: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Mn com diferentes quantidades de atributos.

Quantidades de Atributos						
		7		15		825
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,05	106,10 %	0,08	106,40 %	0,50	92,52 %
M5Rules	0,16	98,57 %	0,08	102,27 %	0,30	95,87 %
LinearRegression	0,15	98,61 %	0,28	99,07 %	0,10	100,09 %

5.4.9 Fe - Ferro

Assumindo Fe como meta os algoritmos aplicados não obtiveram resultados que atingissem o limiar de 0,70, sendo que a SMOReg obteve o maior resultado para coeficientes de correlação de $r = 0,42$ utilizando todos os atributos. A utilização de bases transformadas não melhorou os resultados obtidos pelos algoritmos e alguns casos como a LinearRegression o resultado foi menor do que utilizando bases não transformadas. A Tabela 84 apresenta os resultados obtidos.

Tabela 84: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Fe com diferentes quantidades de atributos

Quantidades de Atributos						
		9		12		825
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,01	103,38 %	0,04	107,59 %	0,42	92,93 %
M5Rules	0,11	101,17 %	0,06	104,55 %	0,37	92,42 %
LinearRegression	0,21	97,68 %	0,14	100,48 %	0,11	105,09 %

5.4.10 Zn - Zinco

Para o Zn os algoritmos não obtiveram resultados de correlação que atingissem o limiar de 0,70, sendo que o algoritmo LinearRegression obteve o maior resultado de $r = 0,40$ utilizando 19 atributos. Correlação esta abaixo do que aquela obtida utilizando bases de dados não transformadas que fazendo uso de 28 atributos obteve um $r = 0,62$. A Tabela 85 apresenta os resultados obtidos pelos algoritmos.

Tabela 85: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de Zn com diferentes quantidades de atributos.

Quantidades de Atributos						
		4		19		825
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,10	100,11 %	0,08	101,77 %	0,24	98,48 %
M5Rules	0,19	97,99 %	0,20	97,66 %	0,20	98,15 %
LinearRegression	0,20	96,67 %	0,40	106,06 %	0,25	99,97 %

5.4.11 B - Boro

Para o B como meta, dentre os resultados obtidos pelos algoritmos, nenhum atingiu o limiar, sendo o maior valor obtido foi pelo algoritmo SMOReg com $r = 0,46$ utilizando 24 atributos. Resultados estes que não obtiveram melhoras em relação aqueles obtidos utilizando bases de dados não transformada. A Tabela 86 apresenta os resultados obtidos.

Tabela 86: Resultados obtidos a partir dos algoritmos de regressão utilizados para estimativa do teor de B com diferentes quantidades de atributos.

Quantidades de Atributos						
		13		24		825
Algoritmo	r	RRSE	r	RRSE	r	RRSE
SMOReg	0,45	89,89 %	0,46	89,76 %	0,50	87,69 %
M5Rules	0,41	91,41 %	0,37	92,98 %	0,37	93,29 %
LinearRegression	0,42	92,38 %	0,43	92,06 %	0,39	95,56 %

6 DISCUSSÃO

Nesta seção são discutidos os resultados apresentados anteriormente.

6.1 Análise Foliar

Os resultados obtidos a partir de análise foliar em espectrofotômetro NIR e MIR são discutidos nesta seção, considerando as bases de dados não transformadas e as que sofreram transformações matemáticas. Correlações acima de 0,70 foram obtidas para a maioria dos nutrientes, exceto para Ca e B. O algoritmo SMOReg considerando a base de dados não transformada obtida por espectroscopia NIR, obteve correlação $r = 0,70$ utilizando os 275 atributos para estimativa do teor de N, resultado este próximo ao encontrado por Santos, Santos e Korndörfer (2012) que utilizou 51 amostras de folhas de soja e teve correlação de 0,76 para o N.

A técnica de espectroscopia NIR apresentou correlações acima de 0,90 fazendo uso do algoritmo LinearRegression para os nutrientes P e K utilizando bases de dados não transformada, sendo que $r = 0,92$ foi obtido pelo algoritmo para o P com 28 atributos, modelo este melhor do que aquele obtido por base transformada com 66 atributos, que foi $r = 0,93$. Para o caso do K a LinearRegression obteve $r = 0,92$ utilizando base de dados transformada e 36 atributos. Resultados estes que obtiveram valores acima do encontrado por Zhai *et al.* (2013), que utilizou-se da técnica de SVM e PLSR para estimar valores de N, P e K em folhas de cereais tais como arroz, milho e soja obtendo resultados para correlações 0,70, 0,72 e 0,70 respectivamente.

Entre os demais nutrientes que obtiveram correlações acima de $r = 0,70$ o algoritmo LinearRegression destacou-se ao obter $r = 0,88$ com 31 atributos para o S utilizando dados não transformados. Para o restante dos atributos o algoritmo SMOReg obteve as maiores correlações sempre utilizando-se de todos os atributos da base de dados, sendo para Mg ($r = 0,70$), Mn ($r = 0,80$) e Fe ($r = 0,83$) quando utilizou-se de base transformada e $r = 0,81$ com base não transformada, sendo Fe com base obtida pelo espectrofotômetro MIR.

Poucos estudos tem abordado a estimativa de Cu e Zn em relação a cultura de soja, neste trabalho o algoritmo SMOReg utilizando-se de base de dados por espectroscopia MIR obteve para Cu correlações $r = 0,87$ com 5.604 atributos a partir de base de dados transformada e $r = 0,85$ com 1.868 atributos utilizando base de dados não transformada, neste caso o modelo é considerado melhor pois utiliza uma quantidade menor de

atributos e apenas uma diferença de 0,02 em relação ao coeficiente r . Para o nutriente Zn o algoritmo SMOReg obteve $r = 0,85$ com base transformada e $r = 0,84$ utilizando-se de base não transformada, mesmo caso do Cu em que o modelo é melhor mesmo com um r menor mas, utiliza menores quantidades de atributos.

O impacto positivo da redução da dimensionalidade na análise foliar pode ser visto em relação aos coeficientes obtidos para P e K. Dentre os demais nutrientes essa redução obteve ganhos pouco significativos em termos de correlação, uma vez que seu acréscimo foi de apenas 0,01, porém o impacto pode ser visto por meio da simplificação dos modelos gerados com um menor número de atributos.

6.2 Análise de Grãos

Os resultados obtidos a partir de análise de grãos em espectrofotômetro NIR e MIR são discutidos nesta seção, considerando as bases de dados não transformadas e as que sofreram transformações matemáticas, bem como os algoritmos de regressão utilizados.

Os atributos para os quais foram obtidas correlações superiores a 0,70, foram: N, P, K, Ca e S. No caso do N, essa correlação satisfatória ($r = 0,72$) foi atingida apenas pelo algoritmo SMOReg e com a base de dados NIR transformada e com todos os atributos. Quando foi utilizada a base de dados não transformada o algoritmo SMOReg obteve $r = 0,71$ com todos os atributos. O algoritmo LinearRegression obteve correlação $r = 0,70$ quando utilizou-se de 12 atributos da base de dados transformada, sendo este considerado o melhor modelo para análise de grãos de soja para N.

Em relação aos demais nutrientes o algoritmo LinearRegression obteve os melhores resultados para coeficientes de correlação a partir de análises de espectroscopia NIR, sendo que para o P foi obtido $r = 0,85$ com 28 atributos da base de dados não transformada. Para o K correlações de $r = 0,77$ com 24 atributos de base transformada e $r = 0,75$ com 15 atributos de base não transformada foram obtidos pela LinearRegression, como também correlação $r = 0,84$ com 24 atributos foi obtido para S.

A estimativa de K e Ca a partir de espectroscopia NIR não tem sido o foco de estudos em relação a grãos de soja. Haughey *et al.* (2013) aplicou a técnica para detectar melanina adulterada em farelos de soja, Oliveira *et al.* (2013) avaliou teores de óleo e proteína em grãos, exemplos de que existe relevância em estimar nutrientes, sendo que correlações acima de 0,70 tem sido obtidas, como exemplo para Ca foi obtido $r = 0,77$

com 24 atributos de base de dados transformada e $r = 0,75$ com 7 atributos de base não transformada. Correlações para K $r = 0,77$ com 24 atributos de base transformada e $r = 0,75$ com 15 atributos de base de dados não transformada.

Os resultados a partir de bases transformadas tem obtido correlações maiores do que aquelas obtidas por bases não transformadas, mas com diferenças de no máximo 0,03 para o r e utilizando-se de uma maior quantidade de atributos, deixando assim o modelo de regressão maior. No trabalho de Proença (2012) ao estimar valores de MO e Argila em dados de solos por espectroscopia e aplicando o mesmo tipo de transformação nas bases mas em atributos específicos, os resultados mostraram-se melhores do que aqueles não transformados.

Com relação a análise de grãos, o impacto positivo da redução da dimensionalidade pode ser visto para P e S. Para os demais nutrientes a redução não obteve ganhos significativos, mas pode ser visto por meio da simplificação dos modelos gerados com um menor número de atributos.

7 CONCLUSÃO

A estimativa de teores de nutrientes em folhas por espectroscopia NIR foi satisfatória para N (0,7), P(0,93), K (0,92), Mg (0,7), S (0,88) e Mn (0,8). Conclui-se que a técnica NIR é mais adequada para os nutrientes P, K e S. A estimativa de teores de nutrientes em folhas por espectroscopia MIR foi satisfatória para Cu (0,85), Fe (0,83) e Zn (0,85), levando a conclusão de que o MIR tem capacidade preditiva similar para estes três nutrientes. Portanto, pode-se concluir que a espectroscopia NIR associada a MIR pode ser usada na predição de teores de nutrientes em folhas de soja, exceto para Ca e B.

A estimativa de teores de nutrientes em grãos por espectroscopia NIR foi satisfatória para N (0,72), P (0,85), K (0,77), Ca (0,77) e S (0,84). Conclui-se assim, que o maior potencial de uso de espectroscopia NIR em grãos está na predição dos nutrientes P e S.

A estimativa de teores de nutrientes em grãos por espectroscopia MIR não foi satisfatória para nenhum nutriente. Portanto, o uso da espectroscopia MIR em estimativa de nutrientes em grãos de soja não é viável de ser utilizada, para tipos de dados como os apresentados nesse trabalho.

A análise dos resultados em relação as técnicas de espectroscopia utilizadas permitem concluir também que a técnica NIR, considerando-se tanto dados foliares como de grãos, estima satisfatoriamente N, P, K, e S, destacando-se na estimativa de P e S.

Em relação a influência da redução de dimensionalidade nos melhores resultados obtidos em folhas de soja, concluiu-se que esta redução levou aos elevados índices de correlação ($r > 0,85$) para P, K e S. Estes resultados foram obtidos com o algoritmo LinearRegression. Nos casos em que os resultados satisfatórios foram obtidos sem o uso da redução de dimensionalidade, o que ocorreu para os nutrientes N, Mg, Cu, Mn, Fe e Zn, tais resultados foram obtidos sempre com o algoritmo SMOReg, o que reforça a eficiência desse algoritmo em conjuntos de dados de alta dimensão.

Em relação ao uso de redução de dimensionalidade para estimativa dos nutrientes em grãos de soja conclui-se que esta redução possibilitou que tais estimativas tenham sido satisfatórias ($r > 0,70$), exceto para N, sempre utilizando o algoritmo LinearRegression.

Concluiu-se que a redução de dimensionalidade associada ao uso do algoritmo LinearRegression foi determinante para a obtenção de índices satisfatórios em grãos, tendo contribuído também para a melhoria dos resultados na estimativa de três nutrientes em folhas.

Finalmente, a metodologia empregada nesse trabalho trouxe efetivamente uma contribuição para viabilizar a análise de folhas e grãos de soja por métodos mais rápidos e não poluentes.

REFERÊNCIAS

- AMORIM, H. V. de. *Manual de Métodos Analíticos para o Controle da Produção de Alcool e Açúcar*. 2^a. ed. [S.l.]: Editora Fermentec/Fealq/Esalq-USP, 1996. 230 p.
- ANAMARI, V. d. A. *et al.* Diagnose nutricional com o uso de sensor óptico ativo em algodoeiro. *R. Bras. Eng. Agríc. Ambiental*, SciELO Brasil, v. 16, n. 11, p. 1159–1165, 2012.
- BORGES, H. B. *Redução de Dimensionalidade em Bases de Dados de Expressão Gênica*. Tese (Doutorado) — Pontifícia Universidade Católica do Paraná, 2006.
- CHAKRABARTI, K.; MEHROTRA, S. Local dimensionality reduction: A new approach to indexing high dimensional spaces. In: CITESEER. *VLDB*. [S.l.], 2000. p. 89–100.
- DAVIES, T. The history of near infrared spectroscopic analysis: Past, present and future. *Analisis*, EDP Sciences, v. 26, n. 4, p. 17–19, 1998.
- FERRARINI, H. *Determinação de teores nutricionais do milho por espectroscopia no infravermelho e calibração multivariada*. Tese (Doutorado) — Universidade Federal do Paraná, 2004.
- FERREIRA, D. *et al.* Comparison and application of near-infrared (nir) and mid-infrared (mir) spectroscopy for determination of quality parameters in soybean samples. *Food Control*, Elsevier, v. 35, n. 1, p. 227–232, 2014.
- FERREIRA, D. S. *Aplicação de espectroscopia no infravermelho e análise multivariada para previsão de parâmetros de qualidade em soja e quinoa=: Application of infrared spectroscopy and multivariate analysis to predict quality parameters in soybean and quinoa*. Tese (Doutorado) — Universidade de Campinas, 2013.
- HART, J. R.; NORRIS, K. H.; GOLUMBIC, C. Determination of the moisture content of seeds by near-infrared spectrophotometry of their methanol extracts. *Cereal Chem*, v. 39, n. 2, p. 94–99, 1962.
- HAUGHEY, S. A. *et al.* The application of near-infrared reflectance spectroscopy (nirs) to detect melamine adulteration of soya bean meal. *Food chemistry*, Elsevier, v. 136, n. 3, p. 1557–1561, 2013.
- HOLMES, G.; HALL, M.; PRANK, E. Generating rule sets from model trees. In: SPRINGER. *Australasian Joint Conference on Artificial Intelligence*. [S.l.], 1999. p. 1–12.
- HOLSHEIMER, M.; SIEBES, A. *Data mining: The search for knowledge in databases*. [S.l.]: Centrum voor Wiskunde en Informatica, 1994.

- IAPAR. *IAPAR - Instituto agrônomo do paraná*. 2015. Acessado em: 2017-03-03. Disponível em: <<http://www.iapar.br/modules/conteudo/conteudo.php?conteudo=863>>.
- JIA, H. *et al.* A feature weighted spectral clustering algorithm based on knowledge entropy. *Journal of Software*, v. 8, n. 5, p. 1101–1108, 2013.
- KANTARDZIC, M. *Data mining: concepts, models, methods, and algorithms*. [S.l.]: John Wiley & Sons, 2003.
- KANTH, K. R. *et al.* Dimensionality reduction for similarity searching in dynamic databases. *Computer Vision and Image Understanding*, Elsevier, v. 75, n. 1, p. 59–72, 1999.
- KINTO, E. A. *Otimização e análise das máquinas de vetores de suporte aplicadas à classificação de documentos*. Tese (Doutorado) — Universidade de São Paulo, 2011.
- KOVALENKO, I. V.; RIPPKE, G. R.; HURBURGH, C. R. Determination of amino acid composition of soybeans (glycine max) by near-infrared spectroscopy. *Journal of agricultural and food chemistry*, ACS Publications, v. 54, n. 10, p. 3485–3491, 2006.
- LEE, J. H.; CHOUNG, M.-G. Nondestructive determination of herbicide-resistant genetically modified soybean seeds using near-infrared reflectance spectroscopy. *Food chemistry*, Elsevier, v. 126, n. 1, p. 368–373, 2011.
- LIMA, K. M. *et al.* Determinação não destrutiva do nitrogênio total em plantas por espectroscopia de reflectância difusa no infravermelho próximo. *Quim. Nova*, SciELO Brasil, v. 31, n. 3, p. 700–703, 2008.
- MALAVOLTA, E.; VITTI, G.; OLIVEIRA, S. Avaliação do estado nutricional das plantas-princípios e aplicações de fosfato. *Avaliação do estado nutricional das plantas: Princípios e aplicações de Fosfato*, 1989.
- MASONI, A.; ERCOLI, L.; MARIOTTI, M. Spectral properties of leaves deficient in iron, sulfur, magnesium, and manganese. *Agronomy journal*, American Society of Agronomy, v. 88, n. 6, p. 937–943, 1996.
- MITRA, P.; MURTHY, C.; PAL, S. K. Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, Institute of Electrical and Electronic Engineers, v. 24, n. 3, p. 301–312, 2002.
- NETO, W. B. *Parâmetros de qualidade de lubrificantes e óleo de oliva através de espectroscopia vibracional, calibração multivariada e seleção de variáveis*. Tese (Doutorado) — Universidade Estadual de Campinas, Campinas, SP, 2005.
- OLIVEIRA, J. I. de *et al.* A pca and spca based procedure to variable selection in agriculture. *Revista Brasileira de Computação Aplicada*, v. 7, n. 1, p. 30–41, 2014.
- OLIVEIRA, M. *et al.* Teores de óleo e proteína em grãos de soja, com diferentes manejos de percevejo, da colheita ao armazenamento, utilizando a espectroscopia no infravermelho próximo (nir). In: *AMERICAS: INTERNATIONAL CONFERENCE ON SOYBEAN UTILIZATION*. [S.l.: s.n.], 2013.

- PASQUINI, C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, SciELO Brasil, v. 14, n. 2, p. 198–219, 2003.
- PLATT, J. C. 12 fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, p. 185–208, 1999.
- PROENÇA, C. A. *Redes neurais artificiais para previsão dos teores de matéria orgânica e argila do solo na região dos campos gerais utilizando espectroscopia de refletância difusa*. Dissertação (Mestrado) — Universidade Estadual de Ponta Grossa. Ponta Grossa, Paraná, 2012.
- SALIBA, P. R. *Desenvolvimento de metodologias analíticas utilizando espectroscopia visível e infravermelho próximo: classificação de açúcares, determinação de brix e acidez total em suco de laranja*. Tese (Doutorado) — Universidade Estadual de Campinas, Campinas, SP, 1999.
- SANTOS, G. A.; PEREIRA, A. B.; KORNDÖRFER, G. H. Uso do sistema de análises por infravermelho próximo (nir) para análises de matéria orgânica e fração argila em solos e teores foliares de silício e nitrogênio em cana-de-açúcar. *Bioscience Journal*, v. 26, n. 1, 2010.
- SANTOS, G. A.; SANTOS, A. P. dos; KORNDÖRFER, G. H. Sistema por infravermelho próximo (nir) para análises de nitrogênio foliar. *Bioscience Journal*, v. 28, n. 1, 2012.
- SANTOS, R. B. L. d. *Método para síntese de técnicas de seleção de atributos disponíveis no software WEKA*. (Trabalho de Conclusão de Curso) — Universidade Estadual de Ponta Grossa, 2016.
- SCHRADER, B. *Infrared and Raman spectroscopy: methods and applications*. [S.l.]: John Wiley & Sons, 2008.
- SHEVADE, S. K. *et al.* Improvements to the smo algorithm for svm regression. *IEEE transactions on neural networks*, IEEE, v. 11, n. 5, p. 1188–1193, 2000.
- SKOOG, D.; HOLLER, F.; NIEMAN, T. *Princípios de Análise Instrumental*, ; trad. Caracelli, I.; Isolani, PC; Santos, RHA; Francisco, RHP. [S.l.]: Bookman: Porto Alegre, 2002.
- SOUZA, D. M.; MADARI, B. E.; GUIMARÃES, F. F. Aplicação de técnicas multivariadas e inteligência artificial na análise de espectros de infravermelho para determinação de matéria orgânica em amostras de solos. *Quim. Nova*, SciELO Brasil, v. 35, n. 9, p. 1738–1745, 2012.
- SUTHA, K.; TAMILSELVI, J. J. A review of feature selection algorithms for data mining techniques. *International Journal on Computer Science and Engineering*, Engg Journals Publications, v. 7, n. 6, p. 63, 2015.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to data mining*. [S.l.]: Ciência Morderna, 2006. 900 p.

- TEIXEIRA, S. *et al.* Data mining algorithms for prediction of soil organic matter and clay based on vis-nir spectroscopy. *International Journal of Agriculture and Forestry*, Scientific & Academic Publishing, v. 4, n. 4, p. 310–316, 2014.
- VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer science & business media, 2013.
- WITTEN, I. H. *et al.* *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016.
- WITTEN, I. H. *et al.* Weka: Practical machine learning tools and techniques with java implementations. 1999.
- WU, D. *et al.* Application of near infrared spectroscopy for the rapid determination of antioxidant activity of bamboo leaf extract. *Food chemistry*, Elsevier, v. 135, n. 4, p. 2147–2156, 2012.
- ZHAI, Y. *et al.* Estimation of nitrogen, phosphorus, and potassium contents in the leaves of different plants using laboratory-based visible and near-infrared reflectance spectroscopy: comparison of partial least-square regression and support vector machine regression methods. *International journal of remote sensing*, Taylor & Francis, v. 34, n. 7, p. 2502–2518, 2013.