UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL

CÂMPUS DE CHAPADÃO DO SUL

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

MAIK OLIVEIRA SILVA

APRENDIZAGEM DE MÁQUINA PARA CLASSIFICAÇÃO DE POPULAÇÕES DE SOJA PARA VARIÁVEIS INDUSTRIAIS COM BASE EM CARACTERES AGRONÔMICOS

CHAPADÃO DO SUL – MS

2022

UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL

CÂMPUS DE CHAPADÃO DO SUL

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

MAIK OLIVEIRA SILVA

APRENDIZAGEM DE MÁQUINA PARA CLASSIFICAÇÃO DE POPULAÇÕES DE SOJA PARA VARIÁVEIS INDUSTRIAIS COM BASE EM CARACTERES AGRONÔMICOS

Orientadora: Profa. Dra. Larissa Pereira Ribeiro Teodoro

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

CHAPADÃO DO SUL – MS

2022

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

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

CERTIFICADO DE APROVAÇÃO

DISCENTE: Maik Oliveira Silva

ORIENTADORA: Dra. Larissa Pereira Ribeiro Teodoro

TÍTULO: Aprendizagem de máquina para classificação de populações de soja para caracteres industriais com base em caracteres agronômicos

AVALIADORES:

Profa. Dra. Larissa Pereira Ribeiro Teodoro

Prof. Dr. Paulo Eduardo Teodoro

Prof. Dr. Rafael Ferreira Barreto

Chapadão do Sul, 28 de fevereiro de 2022.

Documento assinado eletronicamente por **Paulo Eduardo Teodoro**, **Professor do Magisterio Superior**, em 28/02/2022, às 14:50, conforme horário oficial de Mato Grosso do Sul, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rafael Ferreira Barreto**, **Usuário Externo**, em 28/02/2022, às 15:01, conforme horário oficial de Mato Grosso do Sul, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Larissa Pereira Ribeiro**, **Professora do Magistério Superior**, em 28/02/2022, às 15:17, conforme horário oficial de Mato Grosso do Sul, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

LISTA DE FIGURAS

RESUMO

A soja é uma importante alternativa alimentar na nutrição humana, por apresentar alto teor de qualidade proteica. O grande desafio atual dos programas de melhoramento genético é aumentar a produtividade de grãos e o teor de proteína e, pelo menos, manter o teor de óleo. Por isso a importância de se alcançar classificações acuradas de genótipos para teor de óleo e proteína com alta porcentagem de classificação correta. Uma das abordagens promissoras para classificação de variáveis e/ou conjunto de dados complexos é a aprendizagem de máquina (AM). O objetivo foi classificar grupos de genótipos de soja quanto a variáveis industriais com base em caracteres agronômicos utilizando técnicas de AM. O experimento foi conduzido na safra 2019/2020 na Universidade Federal de Mato Grosso do Sul, em Chapadão do Sul e na Universidade Estadual de Mato Grosso do Sul, em Aquidauana**.** Foi utilizado o delineamento de blocos casualizados com duas repetições e 206 populações F_2 de soja. Os caracteres agronômicos avaliados foram: dias para maturação (DM), altura de inserção da primeira vagem (AIV, cm), altura de plantas (AP, cm), número de ramificações (NR), diâmetro da haste principal (DHP, cm), massa de cem grãos (MCG, g) e produtividade de grãos (PROD, kg ha^{-1}). Os modelos testados foram: máquina de vetor suporte (SVM), redes neurais artificiais (ANN), os modelos de árvores de decisão J48 e REPTree (RT) e floresta aleatória (RF). Utilizando técnicas de AM, gerou-se modelos acurados de classificação de variáveis mais complexas que demandam maior tempo para a sua obtenção como óleo e proteínas em soja, com base em caracteres agronômicos, que são de mais fácil mensuração. O RF foi a técnica com melhor desempenho e pode ser utilizada para contribuir com programas de melhoramento genético de soja por meio da classificação de genótipos quanto a caracteres industriais como teor de óleo e proteína.

Palavras-chave: *Glycine max* (L.) Merril. Proteína. Óleo. Random Forest.

ABSTRACT

Soybean is an important food alternative in human nutrition, because it has a high protein quality. The great current challenge of genetic improvement programs is to increase grain yield and protein content and at least maintain oil content. Hence the importance of getting the oil or protein content with a high percentage of correct classification. One of the promising approaches for classifying variables and/or complex data sets is machine learning (AM). The objective was to classify groups of soybean genotypes according to industrial variables based on agronomic characters using AM techniques. The experiment was conducted in the 2019/2020 harvest at the Federal University of Mato Grosso do Sul, in Chapadão do Sul and at the State University of Mato Grosso do Sul, in Aquidauana. A randomized block design with two replications and 206 F_2 soybean populations was used. The agronomic characters evaluated were: days to maturation (DM), height of insertion of the first pod (AIV, cm), plant height (AP, cm), number of branches (NR), diameter of the main stem (DHP, cm), mass of one hundred grains (MCG, g) and grain yield (PROD, kg ha^{-1}). The models tested were: support vector machine (SVM), artificial neural networks (ANN), decision tree models J48 and REPTree (RT) and random forest (RF). Using AM techniques, accurate models were generated for classifying more complex variables that require more time to obtain them as oil and proteins in soybeans, based on agronomic traits, which are easier to measure. RF was the technique with the best performance and can be used to contribute to soybean breeding programs by classifying genotypes for industrial traits such as oil and protein content.

Palavras-chave: *Glycine max* (L.) Merril. Protein. Oil. Random Forest.

SUMÁRIO

INTRODUÇÃO

A soja [*Glycine max* (L.) Merril] é a oleaginosa de maior importância econômica no mundo. A Conab estima que a produção de soja para a safra 2021/22 é de 268,2 milhões de toneladas. O volume representa um crescimento de 5% quando comparada com a temporada passada, o que representa cerca de 12,79 milhões de toneladas a mais a serem colhidas (CONAB, 2022).

A soja é uma importante alternativa alimentar na nutrição humana, pois apresenta alto teor e qualidade proteica, lipídios e vitaminas, e é considerada um alimento funcional que oferece benefício a saúde, além de suas funções nutricionais básicas. A concentração de proteína em grãos de soja varia de 31,7 a 57,9% e a concentração de óleo varia de 8,0 a 25,4%, média de 44,3 e 17,9% (base seca) respectivamente, nos 16.472 acessos do Banco de Germoplasma da Embrapa Soja (EMBRAPA, 2015).

A principal utilização da soja, tanto no Brasil como no resto do mundo, é como matéria-prima para a indústria, que produz óleo e farelo. O farelo, rico em proteína é utilizado principalmente na indústria de rações para aves, suínos e bovinos, sendo esse o principal emprego econômico da soja. Segundo Ramos et al. (2017) atualmente 78% do biodiesel produzido no Brasil vêm do óleo de soja sendo assim a matéria-prima mais utilizada na produção de biodiesel tanto no brasil quanto nos EUA (BIODIESEL, 2015), devido a sua disponibilidade de cultivo em larga escala. Também tem sido utilizado na produção de vernizes, tintas, plásticos, lubrificantes e como insumo para as indústrias de cosméticos e farmacêuticas (ROST, 2013).

Os teores de óleo e proteína dos grãos de soja são governados geneticamente, porém fortemente influenciados por fatores como localização geográfica, condições ambientais e época de plantio, principalmente, durante o período de enchimento dos grãos (BARBOSA et al., 2011).

Um grande desafio atual é aumentar a produtividade e o teor de proteína e, pelo menos, manter o teor de óleo. Por isso a importância de se conseguir o teor de óleo ou proteína, com alta porcentagem de classificação correta, onde os teores não são proporcionais a produção em kg. A soja tem relação negativa entre rendimento de grãos e teor de proteína, conforme constatado por Pípolo et al. (2015). Sendo assim, uma técnica que pode ser adotada para esta estimativa é a aprendizagem de máquina (AM) que pode ser empregada em conjuntos de dados complexos, pois possui característica de aprendizado das relações históricas nos dados, conseguindo assim determinar uma relação entre as variáveis dos genótipos. Essa técnica consegue alterar sua conduta de forma autônoma, com o mínimo de interferência humana durante todo o processo (SAS, 2019).

Neste sentido, algoritmos como máquina de vetor suporte (SVM), redes neurais artificiais (ANN), modelos de árvores de decisão e floresta aleatória (RF), podem ser classificados de acordo com o resultado encontrado em cada ambiente como: aprendizagem não supervisionada, aprendizagem supervisionada e aprendizagem por reforço (RUSSEL & NORVIG, 2013). Eles apresentam a capacidade de utilizar as informações de dados de treinamento (TrainData) para construir modelos que permitem a classificação dos dados de interesse.

A hipótese do estudo é de que é possível classificar genótipos de soja com melhor desempenho para teores de óleo, proteína e cinzas, cuja mensuração é onerosa em termos de tempo e recursos financeiros, utilizando informações de variáveis de mais fácil mensuração e que podem ser coletadas ainda no campo, como as variáveis agronômicas. O objetivo foi identificar a melhor técnica de AM para classificação de grupos (clusters) de populações F_2 de soja formadas com base no seu desempenho para variáveis industriais utilizando caracteres agronômicos como variáveis de entrada nos modelos.

MATERIAL E MÉTODOS

Condução dos experimentos

Os experimentos foram conduzidos na safra 2019/2020, em dois locais. O primeiro experimento foi conduzido na Universidade Federal de Mato Grosso do Sul, câmpus Chapadão do Sul, MS – Brasil (localizada a 18°46" Sul, 52°37"W e altitude média de 810 m). O clima da região é classificado como tropical úmido, com precipitação anual média de 1.850 mm e temperatura anual média de 20,5 ± 7,5 °C. O solo da área experimental foi identificado como Latossolo Vermelho Distrófico (SANTOS et al., 2018) e apresenta as seguintes propriedades químicas: pH (CaCl2) = 4,8; matéria orgânica = 17,6 (g dm-3); P = 5,0 (mg dm-3); H + Al = 5,3; K = 69,0 (mg dm-3); Ca = 1,6 (cmolc dm-3); Mg = 0,5 (cmolc dm-3); capacidade de troca catiônica (CTC) = 7,6 (cmolc dm-3); saturação de bases (V) = 30.0%. O segundo experimento foi realizado na Universidade Estadual de Mato Grosso do Sul, câmpus Aquidauana, MS – Brasil (localizada a 20°27’ Sul, 55°48"W e altitude média de 170 m). O clima da região é classificado como Tropical de Savana, com precipitação média anual de 1200 mm e temperatura média anual de 24ºC. O solo da área foi classificado como Argissolo Vermelho Distrófico de textura arenosa (DONAGEMA et al., 2011), com as seguintes propriedades químicas: pH (CaCl2) = 6,1; matéria orgânica = 19,74 (g dm-3); P = 67,5 (mg dm-3); K = 0,3 (mg dm-3); Ca =5,1 (cmolc dm-3); capacidade de troca catiônica (CTC) = 5,1 (cmolc dm-3); saturação de bases (V) = 45.0%. Em ambos experimentos foi realizada calagem três meses antes da semeadura em cada safra para elevar a saturação de bases a 60%, conforme recomendação de Sousa e Lobato (2004). O calcário utilizado possui poder relativo de neutralização total (PRNT) de 90% e poder de neutralização (PN) de 107%. A porcentagem de CaO e MgO é de 31 e 21%, respectivamente.

Figura 1. Localização da área de estudo em Chapadão do Sul e Aquidauana, Mato Grosso do Sul, Brasil.

Foi utilizado delineamento de blocos casualizados com duas repetições e 206 populações F_2 de soja. As parcelas consistiram de uma linha com três metros de comprimento, com espaçamento de 0,45 m entre fileiras e densidade de 15 plantas por metro. A semeadura ocorreu em outubro de 2019 utilizando preparo convencional do solo (aração e gradagem niveladora). As sementes foram tratadas com fungicida (Piraclotrobina + Metil Tiofanato) e inseticida (Fipronil), na dose de 200 mL do produto comercial para cada 100 kg de sementes para proteção contra o ataque de pragas e fungos de solo. Para a fixação biológica de nitrogênio (FBN), as sementes foram inoculadas com bactérias do gênero *Bradyrhizobium*, sendo utilizada a dose de 200 mL de inoculante líquido concentrado para cada 100 kg de sementes. Os tratos culturais foram realizados de acordo com a necessidade da cultura. A Figura 2 mostra as condições meteorológicas no decorrer do experimento.

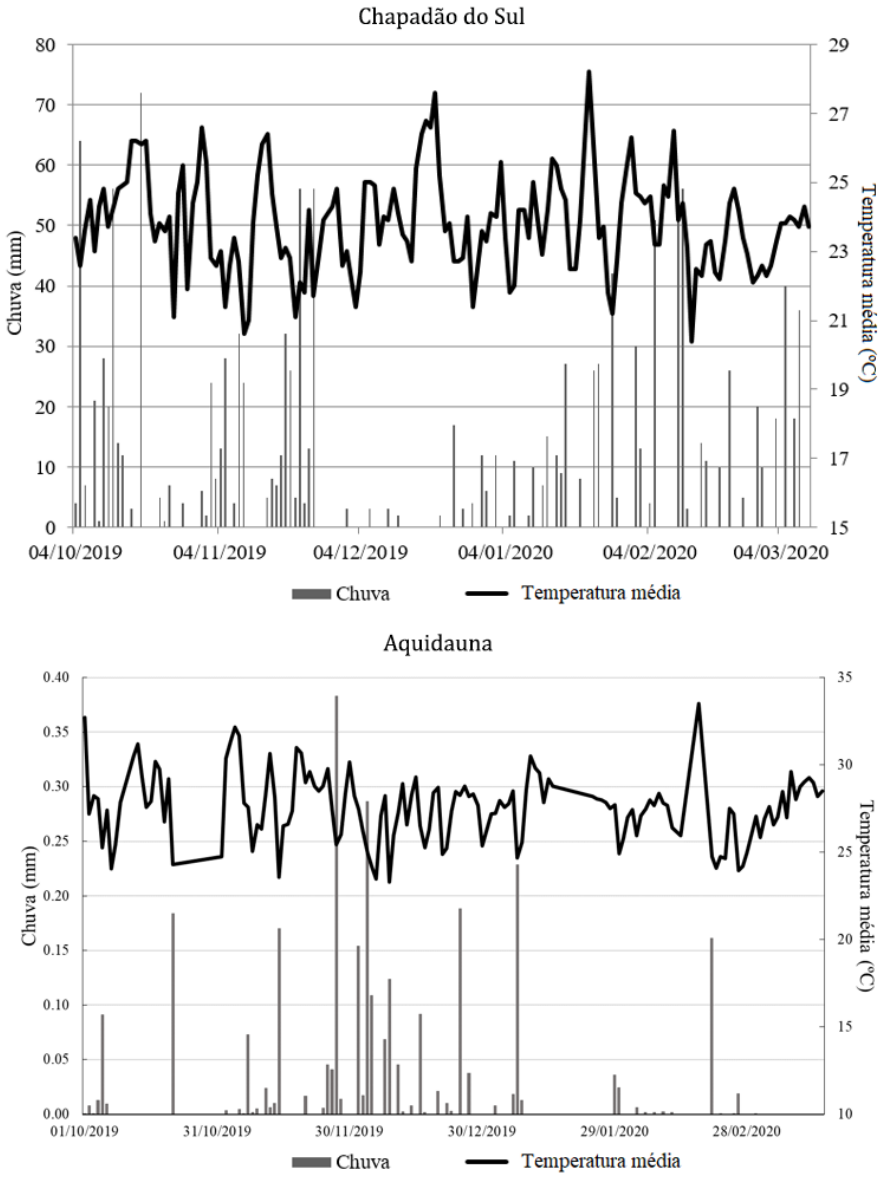

Figura 2. Condições meteorológicas no decorrer da safra 2019/2020 nos munícipios de Chapadão do Sul e Aquidauana.

Variáveis agronômicas

Por ocasião da colheita, foram avaliadas as seguintes variáveis agronômicas: dias para maturação (DM, cm), altura de inserção da primeira vagem (AIV, cm), altura de plantas (AP, cm), número de ramificações (NR), diâmetro da haste principal (DHP), massa de cem grãos (MCG, g) e produtividade de grãos (PROD, kg ha^{-1}). O DM correspondeu aos dias entre emergência e maturação de mais de 50% de plantas em cada unidade experimental. Os

caracteres AIV, AP, NR e DHP foram avaliados em cinco plantas de cada parcela, sendo os dois primeiros avaliados com auxílio de uma fita métrica. Para obtenção da MCG foi retirada uma amostra dos grãos colhidos e a umidade corrigida para 13%. A PROD foi avaliada pela colheita de dois metros centrais de cada parcela e correção para 13% de umidade.

Variáveis industriais

Para determinar os teores de proteína, óleo e cinza (AS) nas populações F_2 de soja, foi utilizada espectroscopia de infravermelho próximo (NIRS) (Metrohm, espectrômetro DS2500, Herisau, Suíça) com alta precisão óptica. As amostras foram homogeneizadas e colocadas em uma cápsula de amostragem. A análise foi baseada na iluminação de uma amostra com radiação de um comprimento de onda específico na região de infravermelho próximo e posterior mensuração da diferença entre a quantidade de energia emitida pelo espectroscópio e refletida pela amostra para o detector (AOAC, 2000). Esta diferença foi medida em várias faixas, criando um espectro para cada amostra. O resultado obtido foi comparado com um conjunto de calibração.

Modelos de Aprendizagem de Máquina - AM

Os modelos de AM testados foram: redes neurais artificiais (ANN), máquina de vetor suporte (SVM), algoritmos de árvores de decisão J48 e REPTree e Random Forest (RF). A técnica convencional de regressão logística (LR) foi utilizada como modelo controle. O SVM executa tarefas de predição construindo hiperplanos em um espaço multidimensional para distinguir as classes diferentes (RAJVANSHI; CHOWDHARY, 2017). A ANN testada consiste em uma camada oculta única formada por um número de neurônios que é igual ao número de atributos, mais o número de classes, todos divididos por 2 (EGMONT-PETERSEN et al., 2002). O modelo de árvore de decisão J48 é uma adaptação do classificador C4.5 que pode ser usado

em problemas de regressão com uma etapa de poda adicional baseada em uma estratégia de redução de erro (SNOUSY et al., 2011). O REPTree utiliza a lógica da árvore de decisão e cria várias árvores em diferentes repetições. Em seguida, seleciona a melhor árvore utilizando o ganho de informação e realiza a poda de redução do erro como critério de divisão (KALMEGH, 2015). O modelo RF é capaz de produzir várias árvores de predição para o mesmo conjunto de dados e usar um esquema de votação entre todas essas árvores aprendidas para predizer novos valores (BELGIU; DRĂGU, 2016). Os seis modelos testados foram executados em uma CPU AMD® PRO A10-8770E R7 com 8 Gb de RAM e todos hiper-parâmetros foram definidos de acordo com a configuração padrão (default) do software Weka (Versão 3.9.4, A Universidade de Waikato, Hamilton, Nova Zelândia).

Para gerar os grupos de genótipos das populações, os dados foram submetidos a análise de componentes principais (PCA). Foi construído um biplot com os dois primeiros componentes principais devido a facilidade de interpretação destes resultados. Neste biplot, foram definidos três clusters (C1, C2, C3) com base no desempenho dos genótipos para as variáveis industriais para utilização do algoritmo k-means, que agrupa tratamentos cujos centroides estão mais próximos até que não ocorra variação significativa na distância mínima de cada observação a cada um dos centroides. Essas análises foram realizadas com auxílio do pacote “ggfortify” do aplicativo gratuito R (R Core Team, 2014). Para as análises de AM, os três clusters formados foram utilizados como variáveis de saída (output) dos modelos, enquanto que as variáveis agronômicas (DM, AIV, AP, NR, DHP, MCG e PROD) foram utilizadas como entrada dos modelos. A classificação dos clusters foi realizada pelos seis modelos de AM em uma validação cruzada estratificada com k-fold=10 e dez repetições (100 execuções para cada modelo).

Análises estatísticas

Para avaliar o desempenho de modelos de classificadores, foram utilizadas as seguintes métricas: porcentagem de classificações corretas e F-score. Essas métricas fazem uso da matriz de confusão, que indica a classificação correta ou incorreta das classes em uso, agrupando os resultados em quatro classes, sendo elas: Falso Negativo (FN), Falso Positivo (FP), Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN).

F-score (também conhecida como F- measure ou Score F1) é uma medida da precisão de um teste, considera tanto a precisão quanto a recordação do teste para calcular a pontuação. O Score F1 pode ser interpretado como uma média harmônica ponderada da precisão e da recordação, em que uma pontuação F1 atinge seu melhor valor em 1 e a pior pontuação em 0. A precisão, também chamada de valor preditivo positivo, é a proporção de resultados positivos que são verdadeiramente positivos. (recall), também chamado de sensibilidade, é a capacidade de um teste para identificar corretamente os resultados positivos para obter a taxa positiva verdadeira (CORNELISSEN, 2019). A Equação (1), demostra como essa métrica é calculada:

$$F1 = \frac{2 * precisao * recall}{precisao + recall} \quad (1)$$

Para avaliação do desempenho dos modelos de AM, boxplots para as médias de classificações corretas (CC, %) e F-score foram apresentados juntamente com os resultados do teste Scott-Knott (SCOTT; KNOTT, 1974) a um nível de 5% de significância. Os boxplots foram executados no software R utilizando os pacotes “ggplot2” e “ExpDes.pt”.

RESULTADOS E DISCUSSÃO

Análise de Componentes Principais – PCA

Com base nos resultados dos componentes principais, foram formados três grupos homogêneos quanto as variáveis industriais avaliadas (Figura 3). Essa análise visa medir o inter-relacionamento entre os tratamentos (genótipos contidos nos clusters) e as variáveis. No segundo quadrante estão a maioria dos genótipos do cluster 1 que estão associados a maior teor de proteínas e cinzas. Já no terceiro quadrante a maioria são os genótipos do cluster 2 que está mais associado a maior teor de proteína e fibra. Os pontos cinzas (cluster 3) estão dispersos nos outros quadrantes não mostrando relação com as variáveis. Segundo Hongyu et al. (2015) para aplicações em diversas áreas do conhecimento o número de componentes utilizados tem sido aquele que acumula 70% ou mais de proporção da variância total. Portanto, o total de 74,4% dos componentes principais 1 e 2 indicam que o gráfico pode ser interpretado de forma precisa.

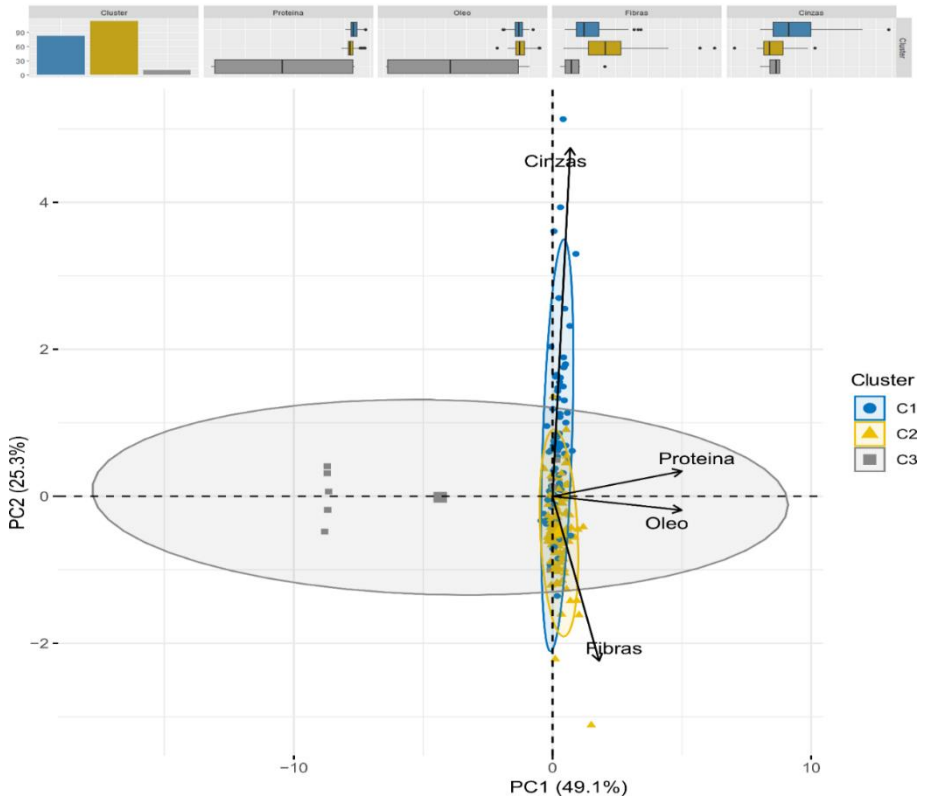

Figura 3. Análise de Componentes principais para os clusters formados pelo algoritmo k-means. Pontos amostrais azuis (círculos) pertencem ao cluster1; pontos amarelos (triângulos) ao cluster 2 e os cinzas (quadrados) ao cluster 3.

Análise de correlação

Segundo o gráfico de correlações e dispersão (Figura 4), a relação dos genótipos dos clusters com as variáveis óleo e proteína (os que estão contidos no cluster 3) mostrou alta variabilidade, já os genótipos dos clusters 1 e 2 apresentaram baixa variabilidade. Os genótipos do C2 mostraram alta variabilidade em fibras e C1 em cinzas. Os genótipos de maior variabilidade são interessantes para seleção de indivíduos para melhoramento genético, onde alguns desses podem ser descartados em futuros estudos de divergência, pois de acordo com Correa e Gonçalves (2012) caracteres que contribuíram com um percentual muito baixo ou não contribuíram para a variabilidade detectada podem ser descartados.

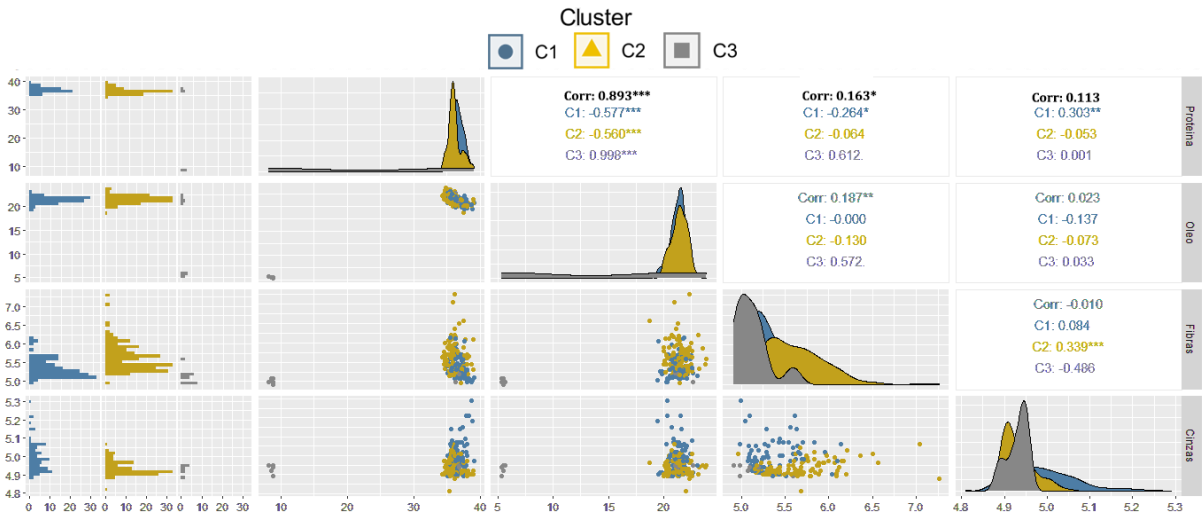

Figura 4. Gráfico de correlações e dispersão entre os clusters e as variáveis proteína, óleo, fibras e cinzas; ***, ** e *: significativo a 0,1%, 1% e 5% de probabilidade respectivamente pelo teste F.

Houve significância entre todos os clusters para a correlação óleo x proteína, porém houve correlação positiva somente para C3. Para a correlação entre fibras x proteínas, apenas houve efeito significativo negativo para C1. Não houve correlação significativa para fibras x óleo considerando os clusters formados. Apenas o C1 apresentou correlação significativa positiva para cinzas x proteína, enquanto que para cinzas x fibras, houve correlação positiva apenas para o cluster C2. Observa-se que os genótipos em C1 e C2 apresentaram correlação negativa para teor de óleo e proteína, corroborando a dificuldade em se obter índices altos das duas variáveis concomitantemente.

Diversos estudos têm reportado a existência de uma associação negativa entre proteína e óleo à medida que se aumenta o teor de proteínas, o teor de óleo é reduzido, e vice-versa (LEE et al., 2019). Porém os genótipos de C3 apresentaram correlação positiva para essas variáveis. Isso demonstra que apesar de diversos outros demonstrarem que os teores de óleo são negativamente correlacionados, existem genótipos em que esta correlação é positiva. Esse resultado é de grande importância para o melhoramento da soja, uma vez que demonstra ser possível a seleção de genótipos com altas médias para ambas as variáveis.

Porcentagem de classificações corretas e F-score

A Figura 5 mostra a porcentagem de classificações corretas (CC) baseadas nas diferentes técnicas de AM utilizada. ANN, REPTree, RF e SVM apresentaram maior porcentagem de classificações corretas enquanto que as técnicas J48 e RL apresentaram médias de CC inferiores. Já para o F-score, as técnicas J48 e RF apresentaram melhor desempenho. Portanto, o RF foi a técnica com melhor desempenho considerando ambas as métricas de porcentagem de classificação.

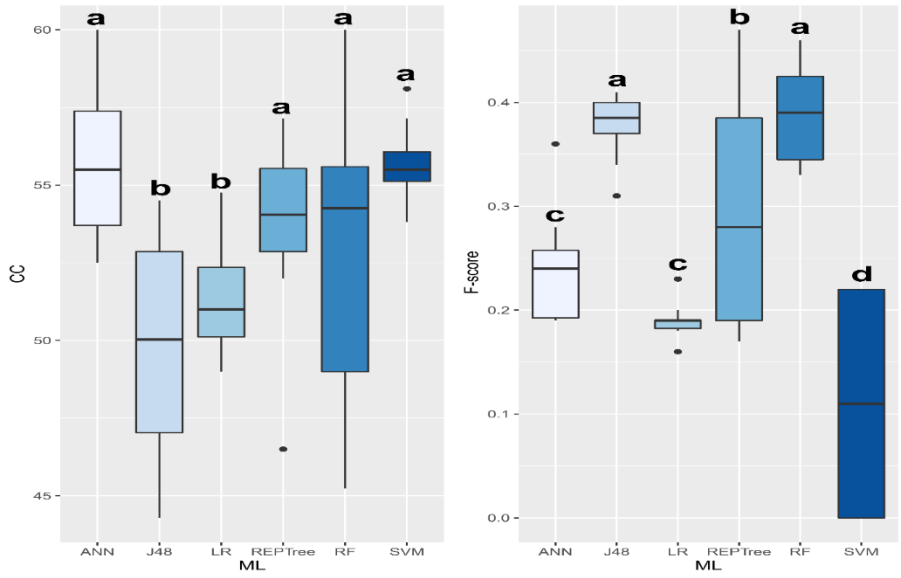

Figura 5. Boxplot para médias de classificações corretas (CC, %) e F-score considerando a análise de regressão logística (LR) e técnicas de AM: redes neurais (ANN), algoritmos de árvores de decisão J48 e REPTree, random forest (RF) e máquina de vetor de suporte (SVM) Grupos de médias com letras iguais não diferem entre si pelo teste de Scott-Knott a 5% de significância.

O RF é comumente utilizado por vários autores que apresentaram resultados satisfatórios em comparação aos demais métodos, especialmente em estudos de classificação utilizando dados espectrais, multiespectrais e hiperespectrais, apresentando a vantagem de não ser tão sensível ao sobre-ajuste dos dados de entrada para a classificação (BELGIU; DRĂGU, 2016). O algoritmo RF é um método que se propõe agrupar variáveis de entrada de dados por meio de várias árvores de decisão que são construídas no momento de treinamento do método dentro do vetor de características de cada arvore, e assim são selecionados aleatoriamente alguns dos atributos que a mesma possui. Uma vez feito isso, calcula-se a entropia apresentada por cada atributo e aquele que possuir a maior é escolhido para separar as classes naquela posição da arvore. A saída do classificador será aquela em que a classe foi retornada como a resposta pela maioria das arvores pertencentes à floresta

(GUEDES, 2014), conforme é possível ver na Figura 6. A principal vantagem do uso do RF é a eliminação de overfitting, problema bastante comum quando se utiliza árvores de decisão, justificando assim o seu desempenho superior à outras técnicas.

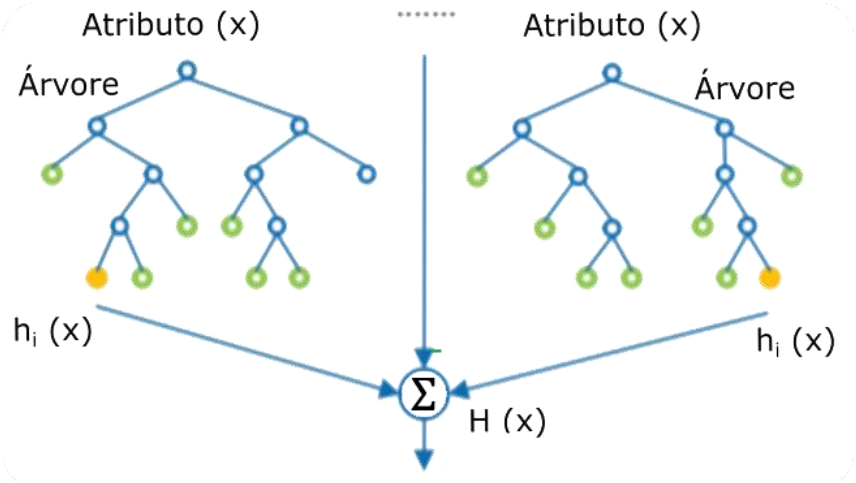

Figura 6. Método de Random Forest. Fonte: Zhao et al. (2017).

Com a eficiência desse método podemos aplicar a AM para a classificação de características mais complexas como teor de óleo e proteína, economizando tempo, mão de obra, recursos e contribuindo para uma melhor tomada de decisão no manejo da cultura da soja.

CONCLUSÃO

O uso de técnicas de aprendizagem de máquina possibilita a classificação acurada de genótipos de soja quanto a variáveis industriais. Com base na porcentagem de classificações corretas e F-score, o Random Forest é a técnica mais eficiente de classificação.

REFERÊNCIAS

AOAC - Association of Official Analytical Chemistry. **Official methods of analysis of the Association of Official Analytical Chemistry,** Washington, p. 4, 2000.

BARBOSA, V. DA S.; PELUZIO, J. M.; AFFÉRRI F. S.; SIQUEIRA, G. B. de. **Comportamento de Cultivares de Soja, em Diferentes Épocas de Semeaduras, visando à Produção de Biocombustível**. Revista Ciência Agronômica, v.42, n.3, p. 742-749, 2011.

BIODIESEL FAQ´s. Jefferson City: National Biodiesel Board, **What is Biodiesel?**. 2015. Disponível em: <http://biodiesel.org/what-is-biodiesel/biodiesel-faqs>. Acesso em: 7 mai. 2015.

BELGIU, M.; DRĂGU, T. L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24–31, 2016.

CONAB - Companhia Nacional do Abastecimento de Grãos. **Acompanhamento da safra brasileira de grãos,** Safra 2021/22, n.5 – Quinto levantamento, Brasília, v. 9, p. 1-101, 2022.

CORNELISSEN, W. G. Detecção automática do início de notas usando a aprendizagem de máquina, p. 1–10, 2019.

CORREA, A. M.; GONÇALVEZ, M. C. **Divergência genética em genótipos de feijão comum cultivados em Mato Grosso do Sul.** Revista Ceres, v. 59, n. 2, p. 206-212, 2012.

DONAGEMA, G.K.; CAMPOS, D. V. B. de; CALDERANO, S. B.; TEIXEIRA, W. G.; VIANA, J. H. M. **Manual de métodos de análise de solos**. 2. ed. Embrapa Solos, Rio de Janeiro, 2011.

EGMONT - PETERSEN, M.; DE RIDDER, D.; HANDELS, H. Image processing with neural networks - a review. **Pattern recognition**, v. 35, n. 10, p. 2279-2301, 2002.

EMBRAPA - **Teores de óleo e proteína em soja: fatores envolvidos e qualidade para a indústria**, comunicado técnico, Londrina, p. 86, 2015.

EMBRAPA - **Sistema brasileiro de classificação de solos**. Centro Nacional de Pesquisa de Solos; Embrapa Solos, Rio de Janeiro, p. 306, 2006.

GUEDES, A. R. M; GUIMARAES, V. L. **Sistemas de Reconhecimento baseado em Random Forest para caracteres de captchas**, Disponível em: <http://www.decom.ufop.br/menotti/rp142/trab/trab1-dp2-artigo.pdf>, 2014.

HONGYU, K. SANDANIELO, V. L. M. JUNIOR, G. J. O. Análise de Componentes Principais: resumo teórico, aplicação e interpretação. **Engineering and Science,** v. 5, p. 89, 2015.

KALMEGH, S. Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. **International Journal of Innovative Science, Engineering & Technology**, v. 2, p. 438–446, 2015.

LEE, S.; VAN, K.; NELSON, R.; LAMANTIA, J.; MCHALE, L. K.; ROUF MIAN, M. A. Genome - wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. **Theoretical and Applied Genetics**, v. 19, p. 21, 2019.

PÍPOLO, A. E.; HUNGRIA, M.; FRANCHINI, J.C.; BALBINOT, A. A. JUNIOR; DEBIASI, H.; MANDARINO, J. M. G. **Teores de óleo e proteína em soja: fatores envolvidos e qualidade para a indústria**. Londrina: Marisa Yuri Horikawa, v. 1, p. 14, 2015.

SANTOS, H.G.; JACOMINE, P. K. T.; ANJOS, L. H. C. dos; OLIVEIRA, V. A. de; LUMBRERAS, J. F.; COELHO, M. R.; ALMEIDA, J. A. de; FILHO, J. C. de A.; OLIVEIRA, J. B. de; CUNHA, T. J. F. **Sistema Brasileiro de Classificação de Solos. Brasília**, 5. ed., 2018.

SAS. **Machine Learning**. Disponível em <https://www.sas.com/pt_br/insights/analytics/machine-learning.html> mai. 2019.

SCOTT, A.J.; KNOTT, M. A. Cluster Analysis Method for Grouping Means in the Analysis of Variance. **Biometrics**, v. 30, n. 3, p. 507–512, 1974

SOUSA, D. M. G.; LOBATO, E. **Cerrado: correção do solo e adubação**. 2. ed, DF: Embrapa Cerrados, v. 2, p. 416, 2004.

RAJVANSHI, N.; CHOWDHARY, K. R. Comparison of SVM and Naïve Bayes Text Classification Algorithms using WEKA. **International Journal of Engineering Research and**, v. 6, p. 09, 2017.

RAMOS, L.P., KOTHE V., CÉSAR-OLIVEIRA M.A.F., MUNIZ-WYPYCH, A.S., NAKAGAKI, S., KRIEGER, N., & CORDEIRO, C. S. Biodiesel: Matérias-primas, tecnologias de produção e propriedades combustíveis. **Revista Virtual** Química 9(1), p. 317-369, 2017.

R CORE TEAM, R*:* **A language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing**, 2014.

ROST, P. T. **Projeto de viabilidade econômico-financeira de implantação de uma indústria para extração de óleo e farelo de soja**. 2013. 99 f. Projeto de Conclusão de Curso de Pós-Graduação em Projetos Empresariais do Curso de Ciências Econômicas do Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, Curitiba, 2013.

RUSSEL, S., NORVIG, P. **Artificial intelligence: a modern approach**. Pearson Education Limited, 2013.

SNOUSY, M. B. A.; EL-DEEB, H. M.; BADRAN, K.; KHLIL, I. A. A. suite of decision tree-based classification algorithms on cancer gene expression data. **Egyptian informatics journal**, v. 12, p. 73–82, 2011.

ZHAO, Y.; MA, X. **Estudo sobre avaliação de crédito de usuários de energia elétrica com base em floresta aleatória (CAC)**, Qingdao, 2017.