Universidade de Brasília
Instituto de Exatas
Departamento de Estatística

Uso de Árvores Aleatórias para Classificação Sensorial de Arroz Cozido

Rafael Lima de Morais

Orientador: Professor Dr. George Freitas von Borries

Brasília

2017

Rafael Lima de Morais

Uso de Árvores Aleatórias para Classificação Sensorial de Arroz Cozido

Relatório final apresentado à disciplina Trabalho de Conclusão de Curso de graduação em Estatística apresentado ao Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Professor Dr. George Freitas von Borries

Universidade de Brasília – UnB

Instituto de Exatas

Departamento de Estatística

Trabalho de Conclusão de Curso de Graduação

Brasília

2017

Resumo

A classificação por floresta aleatória (*random forest*) dispensa suposições paramétricas e possui aplicabilidade em diversos problemas de predição. Foram consideradas as técnicas de floresta aleatória e floresta aleatória para dados desbalanceados. Os resultados foram comparados ao tradicional classificador de regressão logística. As regiões classificadoras foram apresentadas em gráficos de calor da região formada entre as duas primeiras componentes principais, com a intensidade dada pelo valor da probabilidade. Por fim, foi investigado o uso da medida de discrepância fornecida pelo modelo de floresta aleatória, para a determinação de regiões de incerteza de classificação.

A análise foi aplicada em dados de pegajosidade sensorial do arroz cozido. Resultados competitivos foram observados em termos de acurácia do modelo de floresta aleatória quando comparada com a regressão logística. Para o caso específico do arroz de terras altas, do ano de 2014, a floresta aleatória balanceada representou um ganho considerável no desempenho do modelo. Os gráficos de calor apresentados auxiliam na percepção de particularidades do modelo e ajudam no entendimento da construção da floresta aleatória. Finalmente, a barreira construída via valores discrepantes, se mostrou consistente na seleção de observações erroneamente classificadas e quando aplicada à floresta aleatória balanceada dos dados de terras altas, no ano de 2014, resultou em um modelo com apenas um erro de classificação, isso sem custos elevados de não classificação.

Palavras-chave: Árvores de Classificação. Floresta Aleatória. Random Forest. Dados Desbalanceados. Visualização de dados. Regiões de Incerteza. Erro de classificação.

Sumário

Lista de ilustrações

Lista de tabelas

Introdução

Os métodos de classificação estatística visam estabelecer regras relacionando características de observações ao grupo que estas pertencem. Assim, como em um modelo regressivo, é criado a estrutura de variáveis explicativas que devem predizer a resposta, nesse caso, uma variável categórica.

Tradicionalmente, para esse tipo de questão, a técnica mais utilizada é a regressão logística. Em sua forma mais clássica supõe uma distribuição binomial para o padrão de resposta binária. A relação entre as variáveis explicativas e dependente é claramente definida pelos coeficientes regressivos. Por se tratar de um modelo paramétrico, permite inferências com intervalos de confiança e testes de hipóteses. No entanto, o pressuposto de distribuição nem sempre é razoável o que pode limitar seu uso.

Uma segunda abordagem estatística possível é a regressão logística bayesiana. Pelo princípio da verossimilhança, esta técnica é desenvolvida de forma que, não é necessário o conhecimento da distribuição que originou os dados. Em contrapartida, acrescenta a subjetividade na definição de distribuições a *priori*, que em problemas de poucas observações, afeta muito o resultado final.

A técnica estudada neste trabalho pode ser vista como não paramétrica, isto é, não supõe distribuição alguma para a modelagem. Também conhecida no cenário de aprendizado de máquina, os modelos baseados em árvores de decisão utilizam procedimentos intuitivos para modelagem. O classificador utilizado é denominado por floresta aleatória (*random forest*, em inglês) e se baseia na agregação de múltiplas árvores de decisão.

Três grandes interesses relacionados aos modelos são destacáveis: busca por classificadores acurados, visualização de padrões preditivos dos modelos mais complexos e determinação de barreiras consistentes para regiões de incerteza. Primeiramente, consideram-se técnicas baseadas em árvores de decisão, em comparação com modelos de regressão logística. A questão seguinte surgiu da falta de representações gráficas para modelos que combinam árvores. Por fim, observou-se grande relação entre a medida de discrepância construída do modelo de floresta aleatória e o tipo de classificação, correta ou incorreta, que permitiu um paralelo com as barreiras de incerteza apresentadas por Rocha (2017).

Os dados utilizados foram cedidos pela Embrapa Arroz e Feijão (CNPAF-GO), projeto QualiArroz, resultado da parceria dos pesquisadores com o professor George F. von Borries. Este estudo é um dos produtos do grupo de pesquisa do Laboratório de Bioestatística (Labiest)[1]. As análises e abordagem aos dados se baseiam fortemente, nos

[1] Mais informações na página do Labiest <https://labiest.weebly.com> e no diretório de grupos de pesquisa <http://dgp.cnpq.br/dgp/espelhogrupo/8678787481989871>.

estudos de Rios (2015), Oliveira (2015) e Rocha (2017).

Nesse estudo tem-se uma terceira abordagem para a classificação sensorial do arroz, baseado em medidas de perfil viscoamilográfico. Lembrando que a primeira foi a regressão logística politômica apresentada por Rios (2015), seguida por Oliveira (2015) que fez uso da regressão logística bayesiana. Ambas as técnicas obtiveram resultados semelhantes quanto à acurácia. É realizado um estudo comparativo de performance entre os modelos baseados em árvores, com e sem considerar tratamento para grave desbalanceamento, e a técnica de regressão logística.

Para representação gráfica dos modelos de floresta aleatória utilizou-se os chamados gráficos de calor, que aliados à superfície formada pelas duas primeiras componentes principais, permitem a visualização das probabilidades de classificação. Essa metodologia pode ser aplicada a outros modelos de classificação e foi motivada pela falta de apresentações, geralmente associadas a modelos de árvores agregadas.

A barreira de incerteza baseada em valores de discrepância é inspirada pelo estudo de Rocha (2017). Observou-se a consistência desse critério para a identificação de classificações incorretas e boas métricas de avaliação de custo e erros cometidos. Esse tipo de barreira possui a vantagem de propor um critério objetivo para a definição de incertezas, independente das classificações observadas ou preditas.

1 Referencial Teórico

1.1 Análise de Componentes Principais

A análise de componentes principais permite a redução da dimensionalidade de variáveis explicativas. As componentes principais são valores dados por combinações lineares das variáveis, que são definidos de forma que sejam ortogonais entre si juntamente com a maximização da variância dos primeiros componentes. A ortogonalidade garante não correlação linear entre as variáveis, mas não exime a possibilidade de dependência.

Seja a matriz $\boldsymbol{X}_{n\times P}$ a representação das n observações juntamente com as P variáveis. A matriz de variância-covariância (Σ) dos pares de variáveis. Os componentes principais são dados por:

$$CP_i = \mathbf{e}_i^t\mathbf{X} = e_{i1}^t X_1 + \ldots + e_{iP}^t X_P, \quad i = 1, \ldots, P. \tag{1.1}$$

Em que $\mathbf{e}_i$ corresponde ao i-ésimo autovetor da matriz Σ.

Os autovetores e autovalores podem ser definidos pela solução da expressão:

$$|\Sigma - \lambda I| = 0. \tag{1.2}$$

A escala de grandeza das variáveis afeta diretamente o cálculo das componentes principais. São dados pesos maiores às variáveis de maior magnitude, consequentemente não são captadas as reais fontes de variação. Como solução, Johnson e Wichern (2007, p. 436) recomendam que ao invés das variáveis $\boldsymbol{X}$ se utilize uma versão padronizada. As variáveis padronizadas são dadas por:

$$Z_i = \frac{(x_i - \mu_i)}{\sqrt{\sigma_{ii}}}, \quad i = 1, \ldots, P. \tag{1.3}$$

Em que, μ_i representa a média e $\sqrt{\sigma_{ii}}$ o desvio padrão da i-ésima variável.

A matriz de variâncias-covariância (R) associada à variável $\boldsymbol{Z}$ é equivalente a matriz de correlações de $\boldsymbol{X}$. De forma análoga se definem as componentes como:

$$CP_i = \mathbf{e}_i^t\mathbf{Z} = e_{i1}^t Z_1 + \ldots + e_{iP}^t Z_P, \quad i = 1, \ldots, P. \tag{1.4}$$

Em que $\mathbf{e}_i$ corresponde ao i-ésimo autovetor da matriz R.

Os autovetores e autovalores podem ser definidos pela solução da expressão:

$$|R - \lambda I| = 0. \tag{1.5}$$

Por fim, as componentes principais são obtidas pelas combinações lineares que maximizam a variância (JOHNSON; WICHERN, 2007, p. 431), o que permitem redução de fatores explicativos em um modelo de previsão.

1.2 Regressão Logística

De acordo com Hosmer e Lemeshow (2000) a regressão logística consiste de um modelo que relaciona um conjunto de p variáveis independentes $X_1, \ldots, X_P$ a uma variável dependente Y de resposta binária, assumindo valores 0 e 1. O modelo logístico possibilita a estimação da probabilidade de ocorrência de um evento, sendo essa representada por $P_{\boldsymbol{X}}(Y = 1) = \pi(y)$.

O valor $\pi(y)$ assume valores no intervalo de 0 a 1, o que impossibilita um modelo de regressão linear. A função *logito* aplicada ao valor $\pi(y)$ possui suporte de $-\infty$ a ∞, e será utilizada como meio de relacionar as variáveis explicativas com a probabilidade. Essa função que transforma variável resposta é denominada função de ligação, e o *logito* é uma das ligações mais utilizadas. A relação do *logito* e as variáveis explicativas é representada por:

$$\log\left(\frac{\pi(y)}{1-\pi(y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_P X_P. \tag{1.6}$$

Dessa forma, a probabilidade de ocorrência pode ser definida por meio da exponenciação da função *logito*, dada pela Equação 1.7, a seguir,

$$\pi(y) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_P X_P}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_P X_P}}. \tag{1.7}$$

Os parâmetros dos modelos são estimados pelo método da máxima verossimilhança, isto é, uma maximização sobre a função de verossimilhança que é definida por:

$$L = \prod_{i=1}^{l} \pi(y) \prod_{i=l+1}^{n} [1 - \pi(y)]. \tag{1.8}$$

Em que, l é o número de observações que assumiram valor 1 e n o número total de observações.

A estimação deve ser calculada numericamente. Procedimentos de maximização, como o método de *Newton Raphson*, por exemplo, são utilizados para determinar os parâmetros que maximizam a função considerada (AGRESTI, 2003).

1.2.1 Politômica Ordinal

A regressão logística politômica ordinal utilizada, também denominada por modelo *logito* cumulativo, se baseia em chances proporcionais. Hosmer e Lemeshow (2000) apresentam esse modelo como um dos três mais utilizados, para casos ordinais. Para cada categoria, se comparam todas as classes anteriores e equivalente à categoria k com todas as classes acima. A probabilidade da categoria k é dada por:

$$\log\left(\frac{P_{\mathbf{X}}(Y \leq k)}{P_{\mathbf{X}}(Y > k)}\right) = \log\left(\frac{\pi_1(y) + \ldots + \pi_k(y)}{\pi_{k+1}(y) + \ldots + \pi_K(y)}\right) = \beta_{0k} + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_P X_P. \tag{1.9}$$

Em que, β_{0K} representa o intercepto da k-ésima categoria e β_i, para $i = 1, 2, \ldots, P$ indica o efeito causado pela covariável x_i no modelo.

Isolando o termo $P_{\mathbf{X}}(Y \leq k)$ da Equação 1.9 se obtém a probabilidade estimada da variável resposta pertencer à categoria k ou inferior:

$$P_{\mathbf{x}}(y \leq k) = \frac{e^{\beta_{0k}+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_P x_P}}{1 + e^{\beta_{0k}+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_P x_P}}. \tag{1.10}$$

A probabilidade da observação pertencer somente a determinada categoria é dada pela diferença das probabilidades de categorias acumuladas, isto é:

$$P_{\mathbf{x}}(y = k) = P_{\mathbf{x}}(y \leq k) - P_{\mathbf{x}}(y \leq (k-1)), \tag{1.11}$$

no qual, $(k-1)$ está representando a categoria imediatamente abaixo da k-ésima classe.

1.3 Métodos Baseados em Árvores

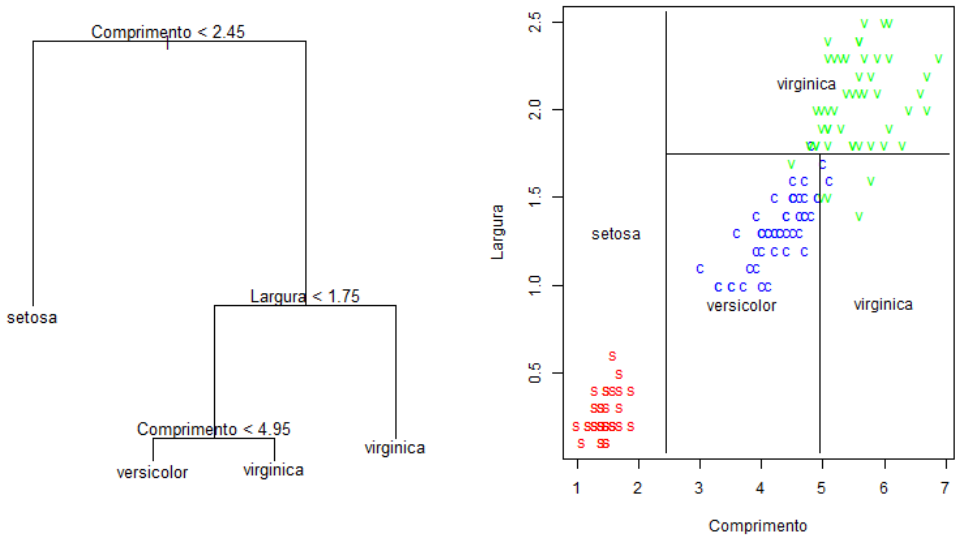
1.3.1 Árvores de Decisão

A técnica de árvores de classificação e regressão (*Classification and Regression Trees - CART*, em inglês) introduzida por Breiman et al. (1984) é de fácil entendimento e serve de base para técnicas mais complexas. Este procedimento tem como objetivo discriminar, em diferentes regiões, observações com base em suas características. Por realizar um particionamento binário recursivo, o modelo final pode ser expresso em uma árvore de decisão. Um excelente ponto de partida para a introdução do algoritmo CART

é o livro de James et al. (2014), o qual, além de introduzir de forma clara os conceitos, também apresenta aplicações no *software* R.

O CART consiste em aplicar a estratificação do espaço, p-dimensional, criado pelas p variáveis preditoras, de forma a gerar regiões disjuntas de classificação. O particionamento hierárquico considera, a cada divisão, todas as variáveis preditoras e através de uma medida de entropia, ou desordem, estipula o critério de divisão que minimiza a variação dentro das regiões. Trata-se de um algoritmo não paramétrico que pode ser utilizado para classificação e regressão, além de admitir preditores quantitativos e qualitativos.

A seguir, Figura 1, exemplifica uma aplicação de árvores de decisão à base de dados Iris (FISHER, 1936). Esses dados são tradicionalmente utilizados na literatura estatística para o estudo de classificação e discriminação.



(a) Árvore de Decisão (b) Dispersão por regiões de classificação

Figura 1 – Exemplo de Árvore de Decisão - Iris

Os dados são compostos por 150 observações de flores, sendo 50 para cada uma das seguintes espécies: “virgínica”, “setosa” e “versicolor”. Cada flor conta com informações de largura e comprimento da pétala e da sépala. O modelo de classificação foi ajustado considerando duas variáveis explicativas (p=2), largura e comprimento da pétala. A primeira imagem mostra o dendrograma do modelo ajustado. Em cada nó da árvore é especificado o critério de partição, por exemplo, na primeira divisão os dados foram divididos em observações cujo comprimento de pétala é superior a 2,45, grupo à direita, contra às que são menores, grupo à esquerda. Desta partição, todos os casos com comprimento inferior foram classificados como “Setosa”. A figura da esquerda permite a visualização do plano formado pelas duas variáveis preditoras, juntamente com o particionamento em sub-regiões

de classificação. A forma da secção hierárquica permite considerar comportamentos não lineares nos dados, para gerar as regiões de classificação.

Seja uma amostra de n observações, de forma que cada indivíduo conta com uma resposta categórica y, assumindo uma das k categorias possíveis, e p variáveis preditoras $\boldsymbol{x} = (x_1, \ldots, x_p)$. O algoritmo automaticamente define as variáveis usadas para partição, bem como seu ponto de corte. Considere, por exemplo, $R_1, R_2, \ldots R_M$, como as M regiões disjuntas, cuja união forma o espaço gerado por $\boldsymbol{x}$, e c_m a resposta categórica associada a cada região. O modelo de classificação pode ser expresso como:

$$\hat{y} = f(\boldsymbol{x}) = \sum_{m=1}^{M} \hat{c}_m I(\boldsymbol{x} \in R_m), \tag{1.12}$$

no qual, $I(.)$ representa a função indicadora, isto é, assume valor igual a 1 quando a condição é satisfeita e 0 caso contrário. O termo R_m representa a m-ésima região de classificação que por ser formado pelas variáveis preditoras e possui p dimensões.

A categoria c_m é atribuída de forma a minimizar o erro de classificação na região R_m. Portanto, define-se $\hat{c}_m$ como sendo o a categoria mais recorrente na m-ésima região, ou seja:

$$\hat{c}_m = \arg\max_k \sum_{\boldsymbol{x}_i \in R_m} I(y_i = k), \tag{1.13}$$

Dessa forma, o algoritmo inicia com toda a amostra para selecionar a variável de corte j e o ponto de partição s. Sendo assim, são geradas duas regiões aninhadas:

$$R_1^i(j,s) = \{\boldsymbol{x} | \boldsymbol{x}_j \leq s\} \quad \text{e} \quad R_2^i(j,s) = \{\boldsymbol{x} | \boldsymbol{x}_j > s\}. \tag{1.14}$$

Em que, o sobrescrito i indica o nível da partição. Em seguida, a variável j e o ponto s são encontrados resolvendo:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1^i(j,s)} Q_1(y_i, c_1) + \min_{c_2} \sum_{x_i \in R_2^i(j,s)} Q_2(y_i, c_2) \right]. \tag{1.15}$$

No qual, $Q_m(y_i, c)$ é uma medida de impureza, isto é, mensura a uniformidade das respostas. Existem diferentes formas de mensurar a impureza, neste trabalho será utilizado o índice de Gini, que é definido por:

$$Q_m(y_i, c) = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) = 1 - \sum_{k=1}^{K} \hat{p}_{mk}^2. \tag{1.16}$$

Em que, $\hat{p}_{mk}$ representa a proporção da categoria k na região R_m. Essa medida é usada como estimativa da probabilidade de cada classe, e é definida por:

$$\hat{p}_{mk} = \frac{1}{n_m} \sum_{\boldsymbol{x}_i \in R_m^p} I(y_i = k). \tag{1.17}$$

A estimativa $\hat{p}_{mk}$ assume valores entre 0 e 1, sendo assim, é possível verificar que $Q_m(y_i, c)$ assume valores pequenos quando todos os $\hat{p}_{mk}$ são próximos de 0 ou 1.

O procedimento apresentado é suficiente para construir um modelo de árvore de classificação com base no CART. Caso não sejam definidos critérios de parada, esse modelo será sobreajustado (*overfitting*) aos dados, isto é, o classificador é capaz de ajustar de forma que nenhuma região estimada possua erro de classificação. Por mais que pareça interessante ter 100% de acerto, é importante destacar que isto se restringe aos dados coletados e não garante baixo erro de previsão, ou seja, classificação de dados não amostrais.

O procedimento de árvores de classificação descrito tornou-se muito popular, primeiramente pela grande interpretabilidade e visualização, do modelo final. As árvores de decisão também possuem a propriedade de lidarem facilmente com preditores qualitativos sem a necessidade de criação de variáveis *dummy*, isto é, indicadoras para as categorias. Por fim, este algoritmo é facilmente otimizado do ponto de vista computacional, podendo ser processado em paralelo.

A estrutura de particionamento hierárquico, adotada pelo CART, é uma estratégia para delimitação dos subespaços, uma vez que é computacionalmente inviável investigar todas as combinações de regiões possíveis. Essa característica torna o procedimento praticável mas, possibilita dois problemas práticos: a solução encontrada pode ser de máximo local, e pequenas alterações nos critérios de partição iniciais podem gerar mudanças drásticas no modelo final, isto é, alta sensibilidade. Outra limitação ao algoritmo proposto por Breiman et al. (1984) é observada quando se utilizam preditores categóricos com muitas classes. Loh e Shih (1997) demonstram que o algoritmo CART tende a priorizar variáveis com muitas categorias como candidatas ao critério de partição, causando assim um viés.

As probabilidades de predição são estimadas pelas proporções obtidas em cada região, o que confere ao modelo uma falta de suavidade característica. Para realizar previsões, isto é, classificar novos dados o CART demonstra baixo poder preditivo, o que dificulta sua generalização.

Por conta da grande interpretabilidade, as árvores de decisão costumam ser usadas de forma exploratória e principalmente de visualização. Seu modelo final possui interpretabilidade direta e intuitiva.

Alguns procedimentos como a *poda* e o tratamento de altas correlação entre os

preditores podem melhorar a performance desse algoritmo para o sobreajuste e alta sensibilidade. Porém, técnicas mais poderosas são recomendadas, para obtenção de um modelo que seja capaz de generalizar o classificador para os dados populacionais. Os próximos modelos a serem abordados pertencem à classe de métodos conjuntos (*ensemble methods*), que segundo Zhou (2012) podem ser divididos em duas categorias: métodos de conjunto sequencial e paralelo. Esses modelos são assim chamados por combinarem muitas árvores de classificação para criar um único classificador, mais robusto e acurado. Nas próximas seções serão descritos dois modelos de conjuntos sequenciais, são eles: *Bagging* e *Random Forest*.

1.3.2 Bagging

O termo *Bagging* vem do acrônimo de *Bootstrap Aggregating* e seu procedimento foi proposto por Breiman (1996b). O mecanismo consiste em combinar duas técnicas chaves, o *bootstrap* e a agregação. Sua intuição parte de combinar vários modelos preditivos, em um único. Pode ser usado para outros algoritmos além do CART, tanto para classificação quanto para regressão mas, o ganho na acurácia só será significativo se os modelos forem instáveis (BREIMAN, 1996b).

Em Breiman (1996a), o autor define como classificador instável, aquele que com pequenas alterações nos dados iniciais produz grandes mudanças no modelo final. Esses modelos altamente sensíveis, usualmente, possuem baixo viés e alta variância, características observadas em modelos sobreajustados.

Seja, $\mathcal{L} = \{(x_i, y_i), i = 1, 2, \ldots, n\}$, conjunto das n observações a serem ajustadas ao modelo. Primeiramente deve se gerar $\mathcal{L}_1^{(b)}, \mathcal{L}_2^{(b)}, \ldots, \mathcal{L}_B^{(b)}$, conjuntos via amostragem *bootstrap*, cada um com n observações. Essa técnica de amostragem consiste em reamostrar os dados com reposição, ou seja, em um conjunto $\mathcal{L}^{(b)}$ se pode observar repetições de uma mesma observação. A técnica de *bootstrap* possui diversas propriedades interessantes e para sua melhor contextualização e detalhamento pode ser verificada em Efron e Tibshirani (1993).

A probabilidade de em n seleções independentes, com reposição, um indivíduo não ser selecionado nenhuma vez é dada por:

$$\begin{aligned} \prod_{i=1}^{n} \tfrac{n-1}{n} &= \left(1 - \tfrac{1}{n}\right)^n \\ &\to e^{-1} \approx 0,367 \end{aligned} \tag{1.18}$$

Essa probabilidade implica que cerca de um terço dos dados não são selecionados. Esse conjunto de observações usualmente é denominado *out-of-bag - OOB* e são muito úteis para estimar o erro de predição do modelo associado a cada $\mathcal{L}^{(b)}$. No contexto de aprendizado

de máquina, as observações OOB são denominadas conjunto de validação, enquanto a amostra *bootstrap* faz o papel do conjunto de treinamento.

Dado os B conjuntos $\mathcal{L}^b$, será ajustado um modelo CART para cada reamostra. A instabilidade de cada classificador é garantida ajustando árvores que não consideram critérios de parada. Portanto, são calculados $\hat{f}_1(\boldsymbol{x}), \hat{f}_2(\boldsymbol{x}), \ldots, \hat{f}_B(\boldsymbol{x})$, tal como na Equação 1.12. O classificador final surge da categoria de maior recorrência da combinação das B árvores ajustadas, ou seja,

$$\hat{f}^{bag}(\boldsymbol{x}) = \arg\max_k \sum_{i=1}^{B} I(\hat{f}_i(\boldsymbol{x}) = k). \tag{1.19}$$

O classificador $\hat{f}^{bag}(\boldsymbol{x})$ considera conjuntamente todos as árvores ajustadas às amostras. Isso contorna o problema de máximo local que ocorre para o CART. Naturalmente, quanto maior o número de árvores consideradas, maior a chance do modelo encontrar o máximo global.

Os estudos de Breiman (1996c), Tibshirani (1996) e Wolpert e Macready (1999) apontam que o erro calculado com as observações OOB é um estimador não enviesado e consistente do erro de predição. Dessa forma, é possível contornar procedimentos onerosos de estimação desse erro como o método, de validação cruzada, *leave-one-out*. Na subseção 2.2.1 da metodologia são abordados os tipos de erro e formas de mensuração. O erro OOB é definido como:

$$err^{oob} = \frac{1}{M} \sum_{i=1}^{M} I(\hat{f}^{bag}(\boldsymbol{x}) \neq y_i^{oob}). \tag{1.20}$$

No qual, M representa o número de observações que não foram selecionadas na amostragem *bootstrap*, $M \approx n/3$. O termo y_i^{oob} representa a classificação real da i-observação. Portanto, o err^{oob} calcula a proporção de vezes em que as classificações foram erroneamente realizadas dado à agregação de M árvores de decisão.

Computacionalmente, esse algorítimo faz uso de comandos recursivos (*loop*) e como padrão, utiliza como critério de parada o número de iterações, amostras *bootstrap*, pré-determinadas.

Seria possível condicionar o número de recursões à estabilização do erro OOB. O mais usual é verificar a convergência do erro após a execução do procedimento. Caso não esteja estável deve-se aumentar as iterações. O pseudocódigo abaixo apresenta a intuição da implementação:

Algoritmo 1: BAGGING

Entrada: $\mathcal{L}$ conjunto de observações da amostra original

Saída: Modelo *Bagging* e erro OOB

para $i = 1$ *até* B **faça**

$\mathcal{L}_i^b = Amostra(\mathcal{L})$ **amostra bootstrap**

$\hat{f}_i(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}|\mathcal{L}_i^b)$ **árvore de classificação sob a amostra** $\mathcal{L}_i^b$

$\hat{f}_i^{bag}(\boldsymbol{x}) = \arg\max_k \sum_{j=1}^{i} I(\hat{f}_i(\boldsymbol{x}) = k)$ **Estimador da *i*-ésima iteração**

$err^{oob} = \frac{1}{i}\sum_{j=1}^{i} err_i^{oob}$ **Erro OOB, média dos *i* erros OOB**

fim

retorna $\hat{f}_B^{bag}(\boldsymbol{x})$ ***e*** err^{oob}

Para cada CART ajustado é obtido $\hat{p}$, como na Equação 1.17. Dado $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_B$ a probabilidade de predição é expressa por.

$$\hat{p}^{bag} = \frac{1}{B}\sum_{j=1}^{B} \hat{p}_i. \tag{1.21}$$

No Apêndice A é apresentado uma visualização da construção de probabilidades dado o número de árvores consideradas.

O procedimento de *bagging* reduz a variância à medida que aumenta o número de árvores consideradas. Ainda assim, o baixo viés dos estimadores instáveis é mantido. Este classificador preserva algumas das propriedades das árvores de decisão como classificar estruturas complexas e não lineares de dados, uso preditores qualitativos e quantitativos. Por conta da agregação, a classificação considera diversos pontos de partida, evitando assim, um possível máximo local. Além de resultar em um procedimento robusto à perturbações e com maior suavidade para probabilidades de classificação.

O alto ganho de acurácia do estimador agregado tem o custo da perda de visualização do modelo. Não é possível representar o classificador *bagging*, em um dendrograma, tal como o CART. Por fim, o procedimento apresentado por Breiman (1996b), resulta em uma grande melhora do classificador instável, mas não possui interpretações adicionais como evidenciação de variáveis mais impactantes para o modelo e medida de observações discrepantes (*outiliers*).

1.3.3 Floresta Aleatória - *Random Forest*

O classificador de floresta aleatória, ou *Random Forest* em inglês, se assemelha ao *bagging* ao considerar diferentes amostras *bootstrap* para cada árvore de classificação. A diferença crucial se deve à seleção aleatória das variáveis preditoras para o CART. Esse

estimador foi apresentado por Breiman (2001), o qual demonstrou propriedades e usos interessantes do método.

Assim como no *bagging*, todas as amostras *bootstrap* são identicamente distribuídas. Isto implica que a esperança da média das B árvores é a mesma que a esperança de cada uma delas, ou seja, o viés do modelo de árvores agregadas será equivalente ao observado em cada árvore. Sendo assim, o modelo de floresta aleatória visa manter o baixo viés de cada classificador individual, à medida que reduz a variância.

O procedimento acrescenta uma etapa ao algoritmo *bagging*. São selecionadas, aleatoriamente em cada iteração, as variáveis que serão utilizadas para o ajuste do CART. No Algoritmo 2 é apresentado pseudocódigo desse modelo de classificação.

Algoritmo 2: Floresta Aleatória - *Random Forest*

Entrada: $\mathcal{L}$ conjunto de observações da amostra original

Saída: Modelo *Random Forest* e erro OOB

para $i = 1$ *até* B **faça**

- $\mathcal{L}_i^b = Amostra(\mathcal{L})$ **amostra bootstrap**
- $\boldsymbol{x^*} = Amostra(\boldsymbol{x})$, **seleciona-se** p^* **variáveis em** $\boldsymbol{x}$**, tal que** $p^* < p$
- $\hat{f}_i(\boldsymbol{x^*}) = \hat{f}(\boldsymbol{x^*}|\mathcal{L}_i^b)$ **árvore de classificação sob a amostra** $\mathcal{L}_i^b$ **dado** $\boldsymbol{x^*}$
- $\hat{f}_i^{rf}(\boldsymbol{x}) = \arg\max_k \sum_{j=1}^{i} I(\hat{f}_i(\boldsymbol{x}) = k)$ **Estimador da** $\boldsymbol{i}$**-ésima iteração**
- $err^{oob} = \frac{1}{i}\sum_{j=1}^{i} err_i^{oob}$ **Erro OOB, média dos** $\boldsymbol{i}$ **erros OOB**

fim

retorna $\hat{f}_B^{rf}(\boldsymbol{x})$ ***e*** err^{oob}

Hastie, Tibshirani e Friedman (2009, p. 588) fornecem uma boa intuição da diferença entre *bagging* e *random forest*, do ponto de vista da variância. Considere, por exemplo, B variáveis aleatórias identicamente distribuídas com variância σ^2 e correlação ρ para cada par. A variância da média é expressa como

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{1.22}$$

O segundo termo diminui à medida que B aumenta, porém, o primeiro permanece inalterado. Portanto, para aumentar a redução da variância, a floresta aleatória gera árvores mais independentes entre si, reduzindo assim a correlação entre os classificadores.

O procedimento de randomização das variáveis preditoras cria um novo parâmetro a ser considerado, o número de características utilizadas em cada iteração. Breiman (2001) observa melhores resultados para $p^* \ll p$, e tipicamente recomenda que se utilize $p^* = \sqrt{p}$ para os modelos de classificação. A probabilidade de predição é calculada da mesma forma

que na Equação 1.21.

Breiman (2001) demonstra, pela lei forte dos grandes números, que à medida que o número de árvores aumenta, isto é, $B \to \infty$, o erro de generalização converge quase certamente para

$$P_{\boldsymbol{X},Y}(m(\boldsymbol{X},Y) \leq 0), \tag{1.23}$$

dado

$$m(\boldsymbol{X},Y) = P_{\Theta}(f(\boldsymbol{X},\Theta) = Y) - \max_{k \neq Y} P_{\Theta}(f(\boldsymbol{X},\Theta) = k). \tag{1.24}$$

A letra Θ representa o conjunto dos vetores aleatórios independentemente distribuídos. O termo $m(\boldsymbol{X},Y)$ expressa a função marginal, à qual, mensura pela média do número de votos, o quanto a classificação correta excede a média do número de votos de qualquer outra classificação. Esse resultado assintótico mostra que à medida que se aumenta o número de árvores, o erro de generalização converge para um limite. Logo, as florestas aleatórias não sobreajustam mesmo que se use infinitas árvores de decisão.

Além de fornecer uma estimativa para o erro de generalização, os dados OOB podem ser usados para o cálculo de importância de variáveis e proximidade de observações. Primeiramente, a medida de importância é de grande utilidade na presença de muitas variáveis preditoras. A importância da variável é dada pela diferença entre a acurácia, ou o índice de Gini, quando se permuta aleatoriamente os n valores da variável e os dados originais, isto mantido fixo os valores das demais variáveis. A medida captura a sensibilidade do modelo para cada preditor, quanto mais sensível, mais importante será a variável para o desempenho do modelo.

As proximidades são expressas em uma matriz de dimensão $n \times n$. A medida de par de observações é dada pela proporção de vezes, em que ambas finalizam, na mesma região de classificação do CART. Breiman (2003) frisa que a medida de proximidade considera a similaridade juntamente com a importância de cada variável. A proximidade varia entre 0, as duas variáveis não foram classificadas nenhuma vez na mesma região, e 1, as duas variáveis sempre foram classificadas na mesma região.

Dada a matriz de proximidades, Breiman (2003) define a medida de discrepância, *oulier*:

$$D_{is}^{0} = \frac{n_s}{\sum_{j=1}^{n} I(y = s) \times prox(\boldsymbol{x}_i, \boldsymbol{x}_j)^2}, \quad i = 1, 2, \ldots, n. \tag{1.25}$$

Em que, n_s é o número de observações pertencentes à categoria s e $prox(\boldsymbol{x}_i, \boldsymbol{x}_j)$ é p valor de

proximidade entre as observações i e j. Em seguida, é definida a discrepância padronizada,

$$D_{is} = \frac{D_{is}^0 - m_s}{MAD^*(D_{is}^0)}, \quad i = 1, 2, \ldots, n. \tag{1.26}$$

No qual, m_s corresponde à mediana de D_{is}^0 e $MAD^*(D_{is}^0)$ é o desvio absoluto mediano (*median absolute deviation - MAD, em inglês*), multiplicado pela constante 1,4826 para que seja estimador consistente da variância. Maronna, Martin e Yohai (2006) apontam a Equação 1.26 como uma padronização robusta a valores extremos, ou seja, assintoticamente normal. Breiman (2003) recomenda a princípio que valores acima de 10 de discrepância sejam considerados como discrepantes, mas dependendo de cada caso, esse valor de corte pode variar.

1.3.3.1 Random Forest para dados desbalanceados

Os classificadores anteriores buscam reduzir o erro geral de classificação pelo voto de maioria. Em um cenário de grande desbalanceamento, o procedimento será incapaz de classificar com precisão categorias menos expressivas. Para o caso de *bagging* e *random forest* uma categoria pouco expressiva provavelmente será pouco amostrada na etapa de *bootstrap*, acarretando em baixo poder preditivo para esses níveis. Visando aumentar a acurácia dessas categorias raras, Chen, Liaw e Breiman (2004) propõem dois procedimentos auxiliares à floresta aleatória, são elas: floresta aleatória balanceada e floresta aleatória ponderada.

A floresta aleatória balanceada se difere do procedimento original na etapa de amostragem *bootstrap*. O balanceamento consiste em estratificar a amostra de forma que todas as categorias possuam o mesmo número de observações selecionadas. Portanto, o nível de maior ocorrência será sub-representado, limitado ao número de observações da classe mais rara.

O segundo modelo atribui penalidades de erro para cada categoria. São endereçados custos maiores às categorias mais raras. Esse procedimento é recomendado para desbalanceamentos mais extremos e incorpora o custo em dois momentos, na função de impureza e no voto de maioria. Na primeira etapa, é calculado o índice de Gini ponderado. Para isso define-se,

$$\hat{p}_{mk} = \sum_{\boldsymbol{x}_i \in R_m^p} \frac{w_k I(y_i = k)}{\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in R_m^p} w_k I(y_i = k)} \tag{1.27}$$

considerando Equação 1.27 calcula-se o índice de gini como na Equação 1.16.

Na segunda etapa, a classificação é dada sobre o critério de voto majoritário ponderado. Novamente o custo irá favorecer as categorias de menores ocorrências.

$$\hat{c}_m = \arg\max_k \sum_{\boldsymbol{x}_i \in R_m} w_k I(y_i = k), \tag{1.28}$$

Todos os demais procedimentos de cálculo do erro OOB, medida de importância e medidas de proximidades, são obtidas de forma análoga ao *random forest* tradicional.

2 Metodologia

2.1 Visualização de Classificações

Quando consideram-se os métodos de agregação de múltiplos classificadores, apresentados na seção 1.3, geralmente não há um recurso de visualização gráfica do modelo ajustado. Para representar como o algoritmo prevê as categorias são utilizados duas ferramentas: componentes principais e gráficos de calor (*heatmaps*).

Primeiramente, as componentes servem para reduzir a dimensionalidade dos dados mantendo o máximo de informação possível. Rios (2015) mostrou que, para alguns subconjuntos de dados do arroz, duas componentes são suficientes para recuperar mais de 90% da variação total. Logo, é utilizado a região gerada pelos escores das duas primeiras componentes para a representação gráfica.

O modelo ajustado fornece probabilidades de classificação para cada categoria, dado as variáveis explicativas. Como a predição é restrita às categorias observadas, mantidas fixas as variáveis explicativas, a soma dessas probabilidades totaliza 1. Considere, por exemplo, uma variável com K níveis de classificação. Logo,

$$p(\hat{y} = 1|\boldsymbol{x}) + p(\hat{y} = 2|\boldsymbol{x}) + \cdots + p(\hat{y} = K|\boldsymbol{x}) = 1. \tag{2.1}$$

Consequentemente, a probabilidade das categorias são complementares entre si. Para um determinável nível i, $p(\hat{y} = i|\boldsymbol{x}) = 1$ se as demais probabilidades forem nulas e $p(\hat{y} = i|\boldsymbol{x}) = 0$ se as demais somarem 1.

Por meio de gráficos de calor (*heatmaps*), é possível representar o comportamento da probabilidade de classificação de cada categoria, sob o espaço formado, entre as duas primeiras componentes. A probabilidade do nível da categoria é representada em termos da intensidade da cor em um espectro que vai de 0 a 1.

O gráfico de calor representa a densidade de probabilidade estimada do modelo para cada categoria. Computacionalmente, é preciso a matriz com os valores de probabilidade de cada ponto do gráfico. Uma forma de simular essa matriz é gerar 100 pontos, igualmente espaçados, variando de acordo com a amplitude observada nos dados, para as duas variáveis explicativas. Em seguida, é calculado a probabilidade de classificação, para cada nível, de todos os pares possíveis das duas sequências, isso gera $100X100 = 1.000$ pontos de probabilidade do espaço formado por duas variáveis preditoras.

Reportando os dados de Fisher (1936), a Figura 2 apresenta os gráficos de calor para os três diferentes tipos de flor, considerando informações de largura e comprimento

da pétala. As probabilidades foram estimadas de um modelo de floresta aleatória com 1000 árvores.

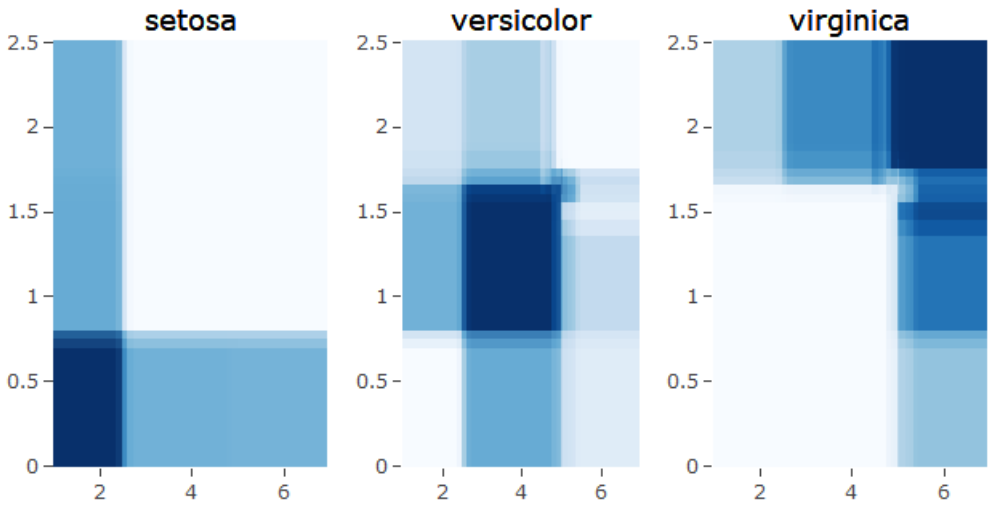


Figura 2 – Exemplo de Gráficos de Calor - Iris

Em comparação com a Figura 1, nota-se que as florestas aleatórias destacam muito bem as áreas retangulares de classificação das espécies de flor. O azul mais escuro, corresponde à probabilidade igual a 1 de classificação. A cor branca, região complementar nos outros gráficos, representa a probabilidade de classificação igual a 0. Dessa forma, as cores de transição correspondem a valores entre 0 e 1. No Código B.3 do Apêndice B são apresentados os comandos que permitem a reprodução dos gráficos acima.

2.2 Métricas de Performance

Há vários critérios que podem ser utilizados para avaliação do desempenho do classificador. Neste trabalho serão considerados procedimentos que podem ser utilizados tanto para a modelagem por regressão logística, quanto por floresta aleatória. O foco das medidas selecionadas é mensurar a acurácia do modelo utilizado, permitindo a comparação nesse quesito.

2.2.1 Mensuração do Erro

A taxa de erro aparente ou taxa de erro de treinamento, é uma medida diretamente obtida do ajuste do modelo. Ela é definida como a proporção de dados classificados incorretamente dentre as observações usadas na modelagem. O erro aparente é facilmente obtido a partir da Tabela 1 que é conhecida como matriz de confusão:

Tabela 1 – Matriz de Confusão

		Classificação Prevista		
	$\boldsymbol{c}$	$y = 0$	$y = 1$	Total
Classificação Real	$y = 0$	n_{0C}	n_{0I}	n_0
	$y = 1$	n_{10}	n_{1C}	n_1

Fonte: Johnson e Wichern (2007, p. 598).

Em que:

- $\boldsymbol{c}$: critério de corte para classificação;
- n_{0C} = número de observações $y = 0$ corretamente classificadas;
- n_{0I} = número de observações $y = 0$ incorretamente classificadas;
- n_{1C} = número de observações $y = 1$ corretamente classificadas;
- n_{1I} = número de observações $y = 1$ incorretamente classificadas.

A diagonal principal da matriz de confusão corresponde à classificações corretas, enquanto os demais valores, nesse caso a diagonal secundária, são classificações incorretas. As matrizes de confusão possuem o mesmo número de linhas e de colunas e podem corresponder a duas ou mais categorias de classificação.

Construída a Tabela 1 das observações de treinamento, o erro aparente é definido como:

$$err^{apa} = \frac{n_{0I} + n_{1I}}{n_0 + n_1}. \tag{2.2}$$

Johnson e Wichern (2007, p. 598) frisam que este procedimento de estimação da taxa de erro é consistente, mas viciado e tende a subestimar a verdadeira taxa de erro. Isto é, a probabilidade esperada de se classificar incorretamente um caso selecionado aleatoriamente. Essa verdadeira taxa de erro também é conhecida como erro de generalização ou de predição.

Um procedimento alternativo é conhecido como método de validação cruzada *jackknife* ou *leave-one-out*. Esse consiste em ajustar o modelo de classificação sem considerar uma observação específica. Esse indivíduo será usado para validar o modelo, ou seja, é realizada a classificação dessa externa. O procedimento é executado até para todas as observações. O erro final é dado pela média de todas as classificações individuais. Trata-se de um procedimento computacionalmente oneroso e é recomendado quando há poucos dados disponíveis.

2.2.2 Curva ROC

A curva ROC (*Receiver Operating Characteristic*) representa graficamente a eficiência preditiva do modelo de classificação binário. A curva é construída ao considerar diferentes limites de discriminação.

Com o auxílio da Tabela 1 são definidas duas medidas importantes de avaliação:

- *Sensibilidade* (Se): proporção de casos positivos reais, sobre o número total de casos positivos. Essa medida está limitada ao intervalo [0,1] e é obtida pela razão n_{1C}/n_1.
- *Especificidade* (Es): proporção de casos negativos reais, sobre o número total de casos negativos. Também é limitada ao intervalo [0,1] e pode ser obtida pela razão n_{0C}/n_0.

A outra medida utilizada para a construção da curva ROC é $[1-(Es)]$ que é a proporção de falsos positivos. Esse valor também é obtido da razão $n_{PF} = n_0$. A curva ROC ótima é alcançada quando os valores de *(Se)* e *(ES)* são iguais a 1, e consequentemente $[1-(Es)]$ igual a 0 para todos os pontos de corte c. Assim, todos as observações seriam classificadas corretamente. Esse caso é representado pela curva vermelha na Figura 3. O mais comum é se observar grandes variações nos valores (Se) e (Es), como representado na curva verde da Figura 3.

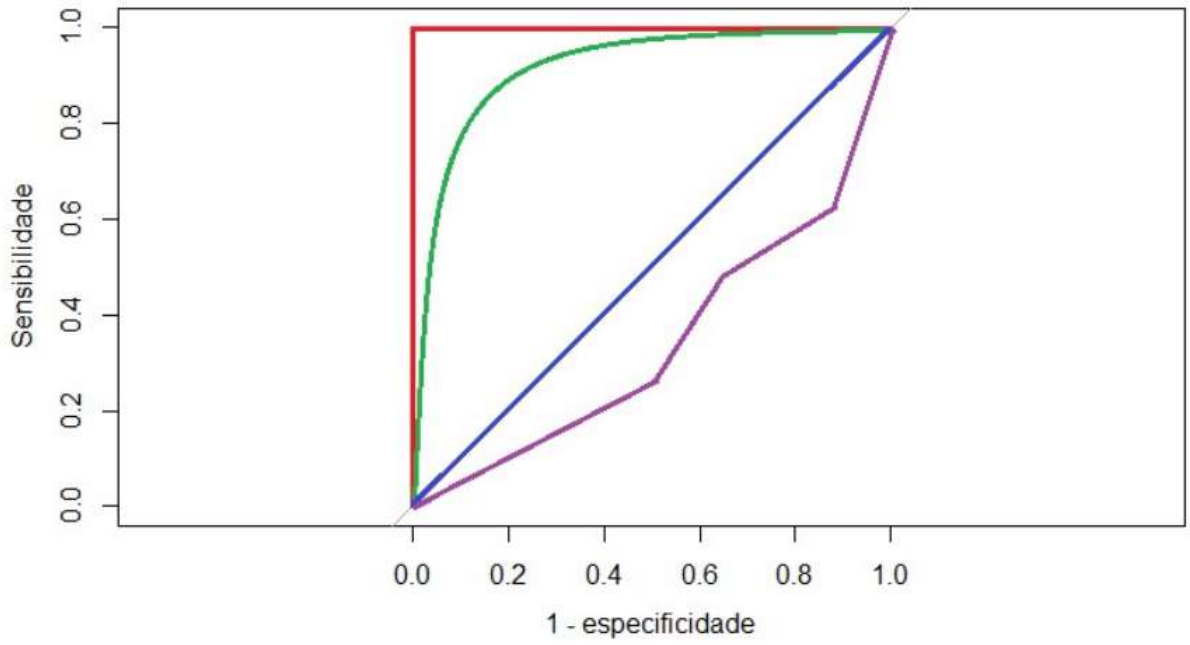


Figura 3 – Exemplos de Curva ROC

Fonte: Rios (2015)

A área sob a curva ROC denominada AUC (*Area Under the Curve*) mensura a discriminação. Isto é a habilidade do modelo de classificar corretamente as duas categorias. Para um modelo com boa discriminação é esperado que o (Se) seja maior do que $[1-(Es)]$, para todos os pontos de corte c.

Com o auxílio da Figura 3, a qualidade do poder de discriminação do modelo pode ser classificada, com base no valor AUC, da seguinte forma:

- discriminação excelente: AUC de 0,9 até 1 - curva vermelha;
- discriminação boa: AUC de 0,8 até 0,9;
- discriminação razoável: AUC de 0,7 até 0,8 - curva verde;
- discriminação ruim: AUC de 0,6 até 0,7;
- discriminação péssima: AUC de 0,5 até 0,6 - curva azul;
- discriminação negativa: AUC de 0,0 até 0,5 - curva roxa.

Um AUC de valor próximo a 0,5 é o equivalente à classificação gerada pelo lançamento de uma moeda honesta, isto é, equivalente à escolha aleatória entre as categorias.

Para o caso de resposta com múltiplas categorias, a curva ROC será calculada para cada categoria, considerando uma contra as demais. A curva ROC que possuir o maior AUC representa a categoria que o modelo está prevendo com maior eficiência, em relação as outras categorias.

2.3 Regiões de Incerteza de Classificação

No estudo de Rocha (2017) são apresentados critérios para definição de diferentes tipos de regiões de incertezas, isto é, locais de transição para a classificação. Essas regiões possuem alto confundimento de categorias acarretando aumento do erro de classificação. Os tipos de barreiras são:

- Barreiras Eliminatórias com base na Dispersão Real (BER);
- Barreiras de Gini (BGI);
- Barreiras de Entropia (BEN);
- Barreiras Eliminatórias com base no Modelo Ajustado (BEM);
- Barreiras Pré-Fixadas (BPF).

Diferentes métricas são calculadas para auxiliar na escolha da melhor barreira de incerteza. Inspirado nessa abordagem foi estudada a informação da medida de discrepância, obtida pela floresta aleatória, sobre a má classificação de uma observação.

2.3.1 Regiões de Incerteza de Classificação via Valores Discrepantes

Por construção da medida de proximidade, altos valores de discrepância são indicativos de indeterminação da classificação predita. Diferente das fronteiras baseadas na modelagem, se observa que esse procedimento também é capaz de captar discrepância em casos de alta probabilidade de classificação, ou seja, longe da região de transição entre as classes.

A medida de discrepância é construída diretamente do modelo de floresta aleatória. Para construir a barreira serão consideradas duas informações de cada observação, o valor $D(y_i)$ de discrepância, e se sua classificação foi correta ou não, definida a seguir:

$$Clas(y_i) = I(y_i = \hat{y}_i), \text{ para } i = 1, 2, \ldots, N. \tag{2.3}$$

Para variável contínua $D(y_i)$, são definidos diferentes pontos de corte e a partir da matriz de confusão, Tabela 2, entre o tipo de classificação e discrepância binária, é investigado o melhor ponto de corte.

Tabela 2 – Matriz de Confusão do tipo de classificação versus discrepâncias

		Discrepância		
	$\boldsymbol{c}$	$D > \boldsymbol{c}$	$D \leq \boldsymbol{c}$	Total
Classificação	*Não*	n_{NI}	n_{NC}	n_N
Correta	*Sim*	n_{SI}	n_{SC}	n_S

A região de incerteza é definida para as observações cujo valor de discrepância supera o ponto de corte $\boldsymbol{c}$. A região de incerteza ideal classificaria todos os casos de classificação incorreta, como casos de incerteza, ao mesmo tempo, que manteria a classificação das observações corretamente classificadas.

Para o cálculo do ponto ótimo de corte é definido:

- $n_{NC}^{\boldsymbol{c}}$: observações incorretamente classificadas fora da região de incerteza considerando o corte $\boldsymbol{c}$;
- $n_{SI}^{\boldsymbol{c}}$: observações corretamente classificadas dentro da região de incerteza considerando o corte $\boldsymbol{c}$.

O ponto de corte ótimo é definido a seguir:

$$\hat{c} = \arg\min_{\boldsymbol{c}} n_{NC}^{\boldsymbol{c}} + n_{SI}^{\boldsymbol{c}} \tag{2.4}$$

Esse corte procura minimizar o erro aparente de classificação atribuindo o mesmo peso para classificações falso positivas e falso negativas. Outras medidas poderiam ser utilizadas, como por exemplo, o índice de Gini ou índice de Gini ponderado. As equações 1.16, 1.17 e 1.27 são facilmente adaptadas para esse exemplo.

A avaliação desse método de definição de barreiras é calculado com base nas seguintes métricas apresentadas em Rocha (2017):

$$\text{Erro de classificação} = \frac{n_{NC}}{n_{NC} + n_{SC}}. \tag{2.5}$$

$$\text{Custo} = \frac{\text{Total de observações na região de incerteza}}{\text{Total de observações}} \tag{2.6}$$

$$\text{Erro Proporcional} = \frac{\text{Total de erros observados nas região de incerteza}}{\text{Total de erros observados originalmente}} \tag{2.7}$$

$$\text{Clas. correta RI} = \frac{\text{Total de observações classificadas corretamente na região de incerteza}}{\text{Total de observações na região de incerteza}} \tag{2.8}$$

Para o caso específico, as métricas acima podem ser calculadas diretamente da Tabela 2:

$$\text{Custo} = \frac{n_{NI} + n_{SI}}{n_N + n_S} \tag{2.9}$$

$$\text{Erro Proporcional} = \frac{n_{NI}}{n_N} \tag{2.10}$$

$$\text{Clas. correta RI} = \frac{n_{SI}}{n_{NI} + n_{SI}} \tag{2.11}$$

Considere a base de dados Iris Fisher (1936) e o modelo de floresta aleatória ajustada na Figura 2. Os valores de discrepância das 150 observações é apresentado na Figura 4. A cor verde indica os casos que foram erroneamente classificados e azul as classificações corretas.

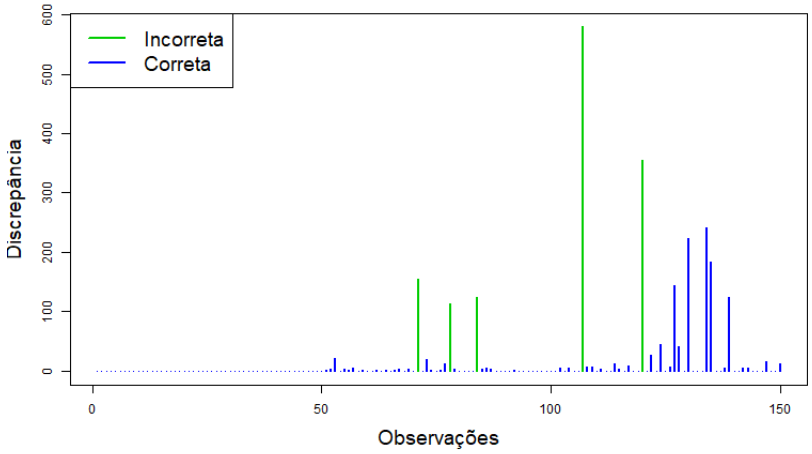


Figura 4 – Valores de Discrepâncias por tipo de classificação - Iris

O ponto de corte que maximiza a Equação 2.4, é de $\hat{c} = 354,24$. A matriz de confusão resultante desse corte é apresentada na Tabela 3.

Tabela 3 – Matriz de Confusão para Discrepâncias - Iris

		Discrepância		
	$\hat{c} = 354,24$	$D \geq \hat{c}$	$D < \hat{c}$	Total
Classificação Correta	*Não*	2	3	5
	Sim	0	145	145

As métricas de regiões de incerteza, apresentadas na Tabela 4, mostram que o uso do valor de discrepância para definir a incerteza não considerou nenhuma categoria corretamente classificada como incerta e ainda identificou 2 dos 5 erros cometidos. O erro de classificação reduziu de 3,33% para 2,03% a um custo de apenas 1,33%. O erro proporcional de 40% indica que a maior parte das observações erradas não foram detectadas com essa barreira.

Tabela 4 – Métricas de Regiões de Incerteza - Iris

Erro de classif.	Custo	Erro prop.	Classif. correta RI
2,03%	1,33%	40,00%	0%

Esses valores são altamente associados ao modelo ajustado. No caso da floresta aleatória para o conjunto de dados Iris o modelo ajusta muito bem e comete poucos erros. Logo, a classificação final com uso de barreiras possui pouco ganho de acurácia e baixo custo.

A curva ROC pode ser utilizada para avaliar o poder de discriminação da medida de discrepância para a classificação correta. A área abaixo da curva obtida é de 95,6%, o que permite afirmar que a a medida de discrepância possui poder de discriminação excelente.

2.3.1.1 Discrepância de novas observações

A matriz de proximidades, utilizada para o cálculo das discrepâncias, é construída juntamente com o processo iterativo da floresta aleatória. Consequentemente é necessário um procedimento de cálculo dessa matriz de proximidade para novas observações. A intuição das barreiras de incerteza é criar áreas cuja classificação é omitida justamente pelo alto grau de confundimento entre categorias.

Uma maneira de estimar a proximidade de uma nova observação, para um modelo de floresta aleatória é dada em duas etapas. Primeiramente, classifica-se a nova observação com base no modelo ajustado previamente. Em seguida, é executado o procedimento de floresta aleatória novamente, considerando a nova observação e sua categoria predita na amostra de treinamento. Deste processo é possível verificar a discrepância da nova observação, dada a categoria predita. Apesar de Breiman (2001) afirmar que seu procedimento, *Random Forest*, é robusto às perturbações, é mais coerente considerar o ponto de corte $\boldsymbol{c}$ definido apenas com base nas observações de treinamento. Resumindo, o segundo modelo ajustado possui a única finalidade de estimar a discrepância da nova observação.

3 Classificação Sensorial de Arroz

3.1 Aplicação

Os dados foram disponibilizados pela Embrapa Arroz e Feijão (CNPAF), projeto QualiArroz, resultado da parceria dos pesquisadores com o professor doutor George F. von Borries. O conjunto de observações conta com 18 variáveis, na qual nove são quantitativas e nove qualitativas. A variável resposta é a análise sensorial da pegajosidade do arroz, ela foi desenhada em uma escala de Likert de 7 níveis. Para permitir comparatibilidade com Rios (2015) e Rocha (2017), são consideradas componentes principais, formadas por variáveis de medidas de perfil viscoamilográfico, para predição da classe sensorial. São 189 observações coletadas no ano de 2013 e 147 no ano de 2014. Cada ano possui dois tipos de arroz, quanto ao cultivo: arroz irrigado (terras baixas) e sequeiro (terras altas).

São reproduzidas condições que permitam a comparação entre as técnicas de floresta aleatória com a regressão logística politômica ordinal, trabalhada por Rios (2015). A comparação é feita por medidas de erro de validação cruzada e curvas ROC. São avaliadas as barreiras de incerteza, por meio de medida de discrepância, por meio das métricas apresentadas por Rocha (2017).

3.2 Pegajosidade de Terras Altas 2014

A escolha dos dados correspondentes ao ano de 2014 se deve ao melhor comportamento para classificação observado por Rios (2015) e Oliveira (2015). O procedimento realizado em detalhes para este conjunto particular pode ser executado para cada um dos demais subconjuntos.

Foram consideradas quatro categorias de pegajosidade do arroz: muito pegajoso (MP), pegajoso (P), levemente solto (LS) e solto (S). Por conta do número de ocorrências a categoria muito pegajoso corresponde à união das observações extremamente pegajosas com as muito pegajosas. O pré-tratamento, realizado neste trabalho, concorda com o procedimento usado em Rios (2015), o que permite a comparação do modelo logístico politômico com o classificador de floresta aleatória.

Na Tabela 5 são apresentadas as frequências de observações de cada categoria de pegajosidade dos dados coletados, juntamente da frequência obtida após a junção das categorias EP e MP.

Tabela 5 – Classificação sensorial de pegajosidade de arroz cozido para as terras altas do ano de 2014

Pegajosidade	Observada	Utilizada
Extremamente Pegajoso	2	-
Muito Pegajoso	11	13
Pegajoso	21	21
Levemente Solto	35	35
Solto	3	3

Fonte: Rios (2015).

Mesmo após a união das categorias extremamente pegajoso e muito pegajoso, ainda se observa um grave desbalanceamento dos dados, visto que a categoria solto corresponde a somente 4% das observações. Vale ressaltar que por conta da agregação, o modelo ajustado não é capaz de prever ocorrências extremamente pegajosas.

As componentes principais foram reproduzidas como em Rios (2015). Foram utilizadas os escores das duas primeiras componentes, as quais juntas explicam mais de 92% da variação total das variáveis viscoamilográficas. Na Tabela 6, é observado, através da carga e da correlação, que as variáveis TAAFIA, TG e FINAL contribuem de forma similar na primeira componente. Para segunda componente destacam-se as variáveis PEAK e BREAKDOWN.

Tabela 6 – Contribuição de cada variável nas duas primeiras componentes principais para arroz de Terras Altas e coeficiente de correlação entre as variáveis e as componentes principais selecionadas

Variáveis	Primeira Componente	Coeficiente de Correlação	Segunda Componente	Coeficiente de Correlação
TAAFIA	0,57	0,9	0,1	0,15
TG	0,51	0,73	-0,31	0,5
PEAK	0,16	0,25	0,66	0,95
BREAKDOWN	-0,28	-0,4	0,6	0,95
FINAL	0,57	0,88	0,3	0,42

Fonte: Oliveira (2015).

Com base na classificação sensorial é possível plotar o gráfico de dispersão dos Escores1 e Escores2, obtido das componentes principais, ilustrado na Figura 5. As diferentes categorias estão em maior ou menor medida agrupadas, com exceção da categoria solto (S), é possível indicar claramente regiões com a prevalência de determinada categoria. Próximo

do ponto (-1;-1) há uma confusão entre as categorias MP, P e LS, o que indica uma área de transição de categorias. Outras regiões de confundimento estão próximas a (0;0) e a (2;1).

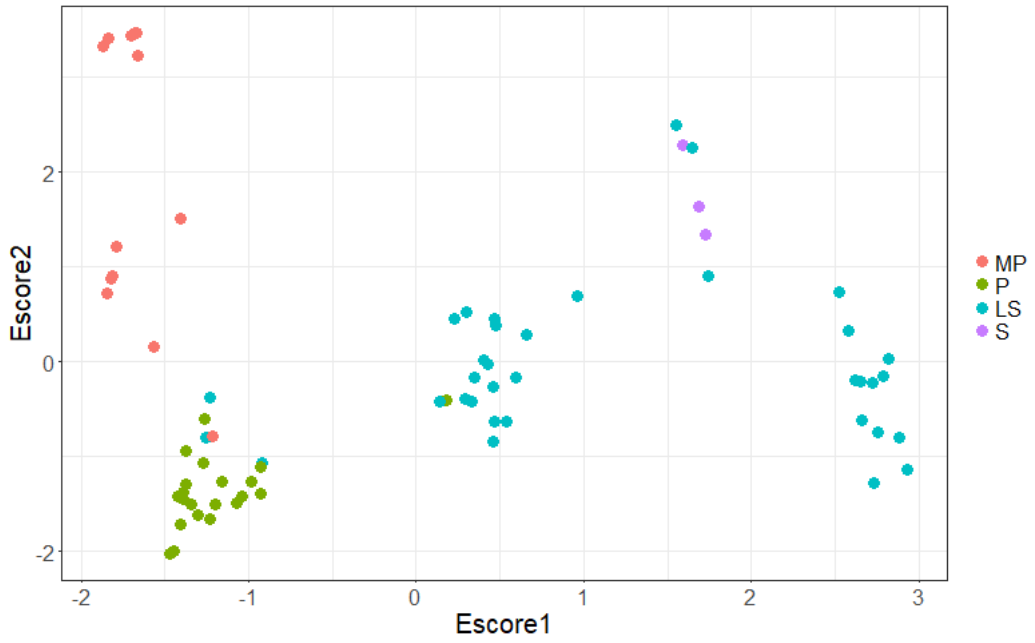


Figura 5 – Dispersão da Classificação Sensorial com base nos Escores1 e Escores2

Sob a perspectiva de Rocha (2017) é possível apontar na Figura 5 a presença de regiões de incerteza com base na classificação do avaliador.

3.2.1 Floresta Aleatória - Random Forest

A floresta aleatória foi ajustada às 72 observações de arroz de terras altas para o ano de 2014. Foram consideradas 1000 árvores de decisão, no qual, para cada, foi escolhida aleatoriamente os escores de uma das duas primeiras componentes principais, como variável explicativa.

O cálculo do modelo no *software* R é rapidamente obtido, ainda que se considere valores maiores, para o número de iterações. Isto se deve ao volume de dados não ser muito grande e ao procedimento computacional, em paralelo, utilizado pelo algoritmo.

A Figura 6 apresenta o erro OOB geral e para cada uma das 4 categorias de pegajosidade, conforme o número de árvores. Com exceção da categoria solto, o erro estabiliza-se rapidamente para as demais categorias. Esse erro está associado ao erro de predição, logo, se espera essas taxas de erro para predição de novos dados de uma mesma distribuição de pegajosidade de arroz.

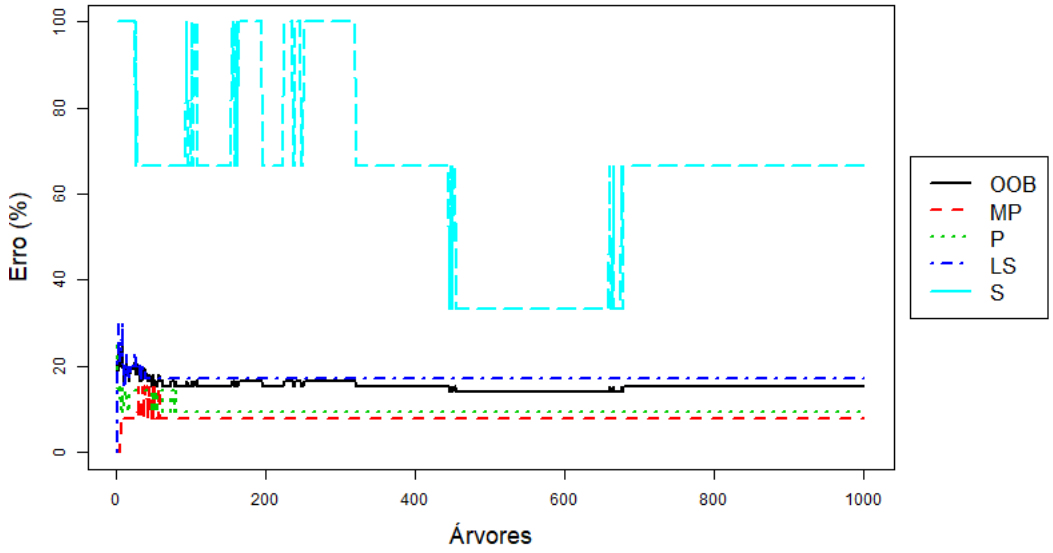


Figura 6 – Erro OOB por Número de Árvores

Os valores finais de erro OOB são apresentados, juntamente com a matriz de confusão, na Tabela 7. De 72 observações 11 foram erroneamente classificadas, com base nos dados OOB, o que corresponde a 15,28% de erro OOB. A categoria melhor ajustada foi o muito pegajoso seguido pela pegajoso, ambas com erros abaixo de 10%. A categoria levemente solta apresentou taxa de erro de 17,48%, mas cometeu erros em todas as categorias. Por fim, a categoria solto teve apenas 1 observação corretamente classificada, dentre as 3 observadas.

Tabela 7 – Matriz de Confusão para a previsão OOB do modelo de floresta aleatória

		Classificação Prevista				Erro
		MP	P	LS	S	OBB
Classificação Real	MP	12	0	1	0	7,7%
	P	0	19	2	0	9,5%
	LS	1	3	29	2	17,1%
	S	0	0	2	1	66,7%

A seguir, Figura 7, é apresentado novamente o gráfico de dispersão das observações pelos escores das duas primeiras componentes principais. Dessa vez, além cores dos pontos indicarem a categorial real da observação, é apresentado a informação se a classificação foi correta ou não. Os pontos vazios são os casos em que a predição se distinguiu da classe verdadeira, o que caracteriza um erro de classificação.

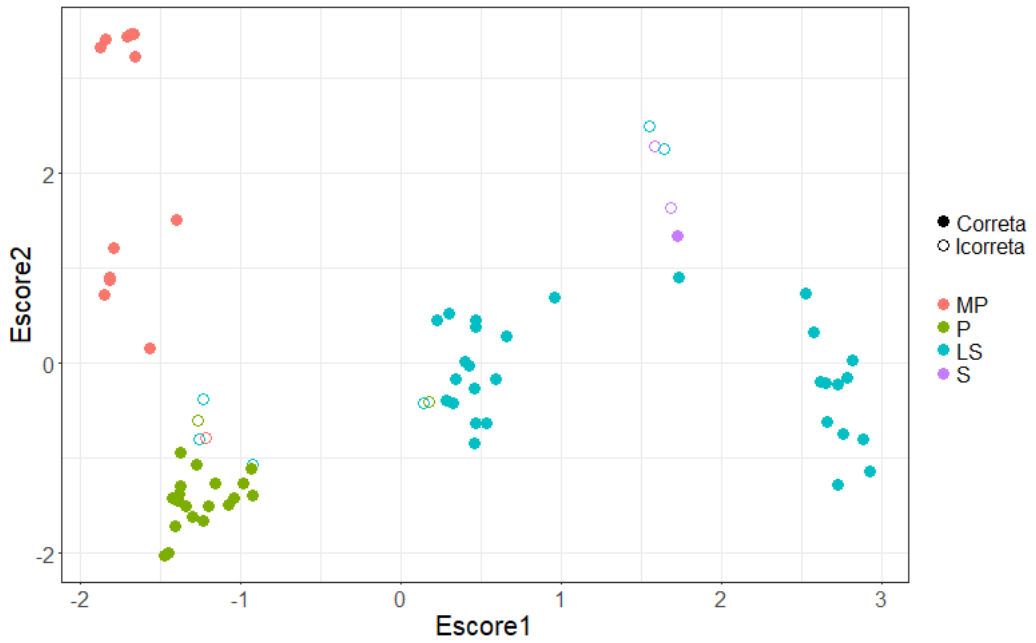


Figura 7 – Dispersão da Classificação Sensorial e Predita com base nos Escores1 e Escores2 para o modelo de floresta aleatória

Como o esperado, o modelo apresenta dificuldades de classificar regiões de confundimento entre as classes. Outra representação gráfica é possível por meio das probabilidades de classificação, expressas em gráficos de calor, apresentado na Figura 8.

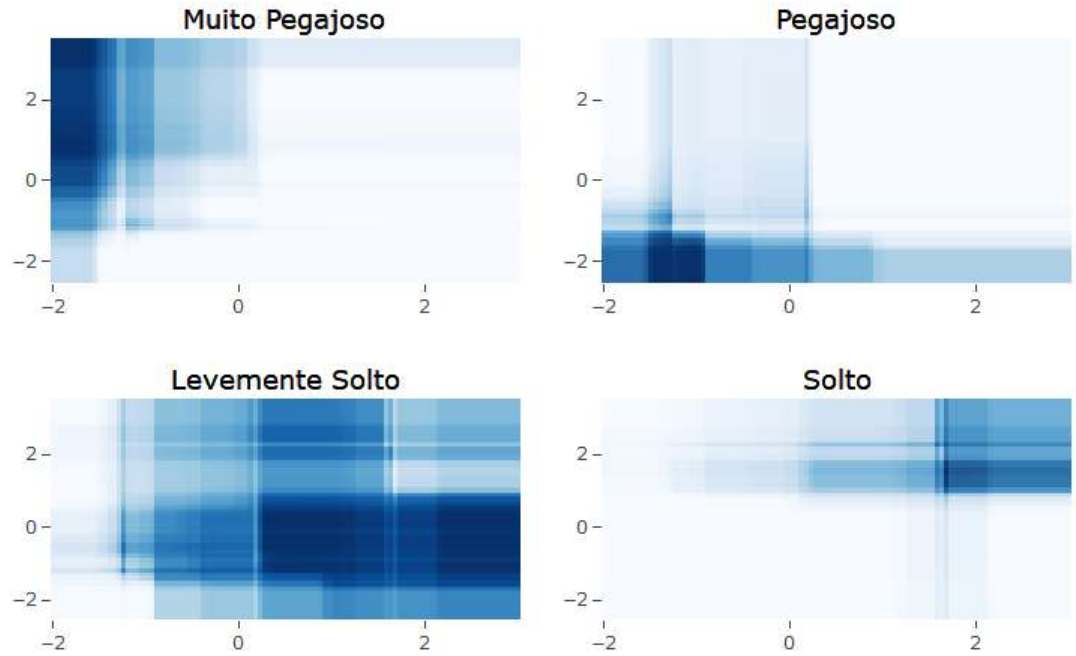


Figura 8 – Gráficos de Calor para Probabilidade de Classificação

As cores mais escuras indicam probabilidades mais altas. Os quatro gráficos de

calor são complementares, somam probabilidade igual a um, o que permite identificar o grau de certeza que o modelo classifica cada observação. Análoga às regiões de incerteza apresentadas por Rios (2015) e Rocha (2017), as áreas cuja cor é semelhante, para mais de uma categoria, indicam barreiras de transição.

A Figura 8 também evidencia a forma de classificação não linear do classificador de floresta aleatória. As regiões de probabilidades estimadas se assemelham à dispersão apresentada na Figura 7. As regiões de probabilidades apresentam formatos retangulares por considerar a região conjunta de 1000 árvores de decisão, cujas probabilidades de classificação também possuem este formato, mas com bem menos suavidade, para as transições. Diferente dos modelos de regressão que estimam uma função contínua, a função de probabilidade estimada pelo modelo de floresta aleatória é dada em saltos de probabilidade.

3.2.1.1 Comparação com Regressão Logística

A comparação entre os dois modelos é dada quanto à acurácia da classificação. São utilizadas técnicas de validação cruzada, para estimar o erro de generalização, e curva ROC como segunda métrica comparativa e de representação gráfica, da classificação de cada nível da variável de pegajosidade do arroz. Para ambos os modelos foram utilizados os escores dos, dois primeiros, componentes principais.

As Tabelas 8 e 9 apresentam a matriz de confusão para os modelos de floresta aleatória e regressão logística politômica ordinal. Foram considerados classificações por meio de validação cruzada.

Tabela 8 – Matriz de Confusão para o modelo de Regressão Logística por Validação Cruzada

		Classificação Prevista				Erro de
		MP	P	LS	S	Classificação
Classificação Real	MP	10	3	0	0	23,0%
	P	0	20	1	0	4,7%
	LS	0	3	32	0	8,5%
	S	0	0	3	0	100%

Tabela 9 – Matriz de Confusão para o modelo de Árvores Aleatórias por Validação Cruzada

		Classificação Prevista				Erro de
		MP	P	LS	S	Classificação
Classificação Real	MP	12	0	1	0	7,6%
	P	0	19	2	0	9,5%
	LS	1	3	29	2	17,1%
	S	0	0	2	1	66,7%

A matriz de confusão para o modelo de floresta aleatória é idêntica à matriz obtida via dados OOB, o que corrobora com a premissa de que o erro OOB é um estimador consistente do erro de generalização.

O modelo de regressão logística apresenta menor erro de classificação, 13,89% enquanto o de floresta aleatória teve 15,28% de erro. Na prática o modelo de árvores classifica erroneamente uma observação a mais que a regressão. Destaca-se que apesar de acurácias bem próximas, os modelos apresentam padrões de classificação distintos. Por considerar a natureza ordinal dos dados, o modelo de regressão logística não comete erros em categorias não vizinhas, isto é, todos os erros de classificação se deram em categorias mais ou menos intensas. A floresta aleatória não distingue resposta categórica de categórica ordinal, o que faz com que o modelo não penalize classificações mais distantes.

A característica linear do modelo logística parece impossibilitar a classificação da categoria mais extrema e de menor ocorrência. Como apresentado na Figura 5, a categoria solto se confunde de forma não linearmente separável com a categoria de arroz levemente solto. Já a floresta aleatória consegue captar, com limitações pelo número de observações, a estrutura não linear que separa essas categorias, Figura 8.

Os erros de classificação do modelo de regressão logístico foram menores para níveis com maiores números de ocorrências. Em contrapartida, o modelo de floresta aleatória apresenta erros mais balanceados.

As Figuras 9 e 10 apresentam as curvas ROC para cada categoria predita por meio de validação cruzada, considerando os modelos de floresta aleatória e regressão logística.

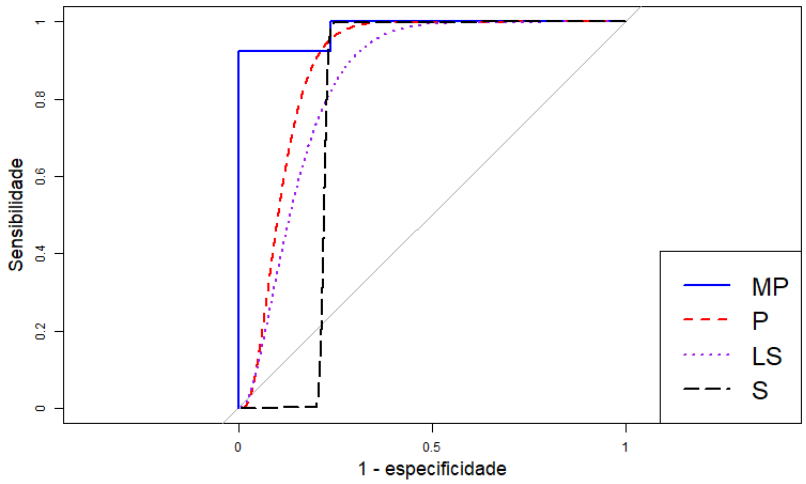


Figura 9 – Curva de classificação ROC da avaliação sensorial de pegajosidade para o ano de 2014, prevista por meio do modelo de regressão logística

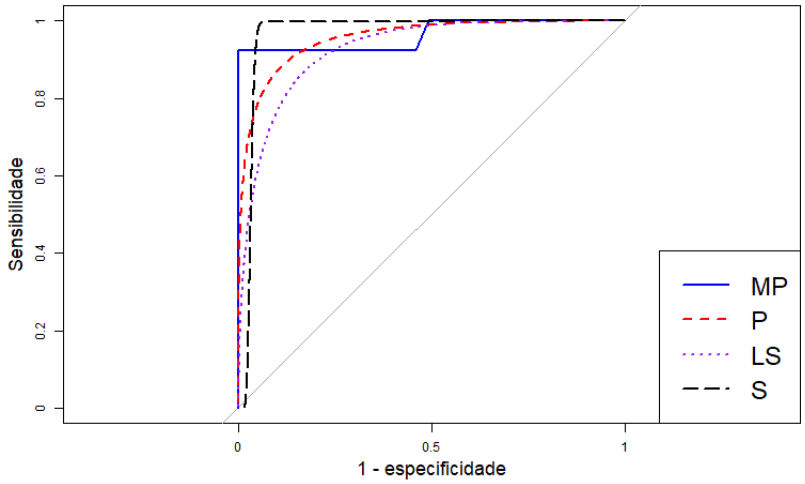


Figura 10 – Curva de classificação ROC da avaliação sensorial de pegajosidade para o ano de 2014, prevista por meio do modelo de floresta aleatória

A Tabela 10 apresenta as áreas abaixo da curva para cada categoria ajustada pelos modelos.

Tabela 10 – Área sob a curva ROC

	Pegajosidade			
Modelo	MP	P	LS	S
Reg. Logística	0,9817	0,8858	0,8465	0,7789
Floresta Alea.	0,9634	0,9555	0,9264	0,9670

Apenas para classe meio pegajoso a regressão logística apresentou maior poder de discriminação, sendo superada em todas as demais classes.

Vale ressaltar que o resultado apresentado pelas curvas ROC não é necessariamente contraditório às medidas de acurácia da validação cruzada. Os valores de sensibilidade e especificidade são calculados sob as tabelas de confusão, tal como, na Tabela 1. A curva obtida pela variação do ponto de corte $\boldsymbol{c}$, mede o poder de discriminação da probabilidade estimada para a categoria observada.

3.2.1.2 Barreira de Incerteza

Dado o modelo de floresta aleatória ajustado, são calculados valores de discrepância para cada observação de arroz cozido. A Figura 11 apresenta aos valores encontrados e distingue pela cor se a classificação foi correta, azul, ou incorreta, verde.

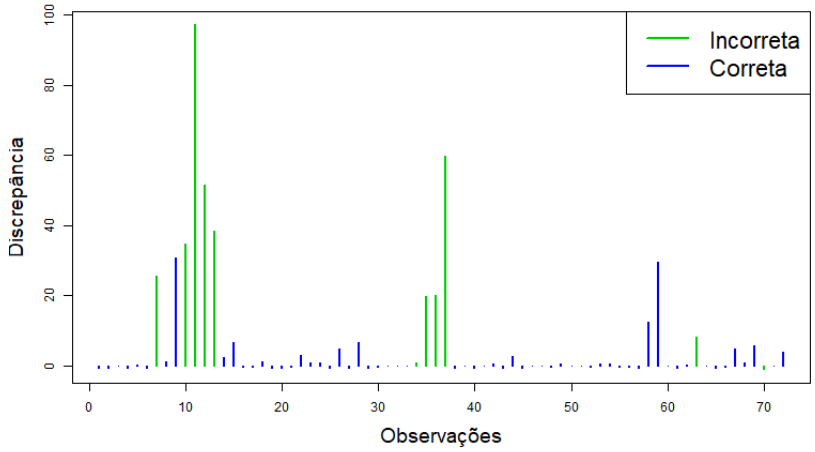


Figura 11 – Valores de discrepância por tipo de classificação

O ponto de corte que maximiza a Equação 2.4, é de $\hat{c} = 13,93$. A matriz de confusão resultante desse corte é apresentada na Tabela 11.

Tabela 11 – Matriz de Confusão para Discrepâncias

		Discrepância		
	$\hat{c} = 13,93$	$D \geq \hat{c}$	$D < \hat{c}$	Total
Classificação	*Não*	8	3	11
Correta	*Sim*	2	59	61

A região de incerteza encontrada inclui 10 observações, das quais 8 foram incorretamente classificadas e 2 corretamente. As métricas de regiões de incerteza, são exibidas na Tabela 12.

Tabela 12 – Métricas de Regiões de Incerteza

Erro de classif.	Custo	Erro prop.	Classif. correta RI	AUC
4,84%	13,89%	72,73%	20,00%	87,03%

O critério de barreira, baseado na discrepância, apresenta custo de 13,89%, ou seja, 11 observações em 72 não serão classificadas. O erro proporcional foi de 72,73% indicando consistência da discrepância na identificação do erro. Apenas 20% das observações classificadas corretamente foram incluídas, na região de incerteza.

A área abaixo da curva ROC foi igual a 87,03%, concorda com o alto valor do erro proporcional, e o baixo valor da classificação correta na região de incerteza, e permite classificar como boa a discriminação entre classificações corretas e incorretas por meio da medida de discrepância.

A seguir, a Tabela 13, apresenta a classificação prevista, após a retirada das observações que pertencem à região de incerteza, versus a real categoria do arroz cozido.

Tabela 13 – Matriz de Confusão para o modelo de Regressão Logística por Validação Cruzada

		Classificação Prevista				Erro de
		MP	P	LS	S	Classificação
Classificação Real	MP	11	0	0	0	0%
	P	0	18	0	0	0%
	LS	0	1	29	0	3,3%
	S	0	0	2	1	33,3%

Após o filtro da região de incerteza, não se observa classificações distantes da vizinhança da categoria. Em outras palavras, os erros que restaram se deram nas categorias adjacentes à classificação real. Isso evidência a consistência da medida de discrepância que resulta em "erros menos graves", visto que, a variável de pegajosidade sensorial é dada em escala ordinal.

Considerando o modelo de floresta aleatória sob as duas primeiras componentes principais, esse é o modelo mais acurado. Por conta do poder de generalização desse procedimento, espera-se que para novas observações de arroz cozido, geradas sob as mesmas condições, o modelo consiga prever bem o comportamento de pegajosidade sensorial.

3.2.2 Floresta Aleatória para Dados Desbalanceados

Nesta seção são detalhados os resultados do modelo de floresta aleatória balanceada. A escolha é justificada por sua performance, que supera a acurácia do modelo de floresta aleatória ponderada.

O modelo exige que se declare o tamanho da amostra a ser selecionada via *bootsrap* para as categorias. A consistência dos resultados para amostra OOB são mantidas se a amostra *bootstrap* for do mesmo tamanho que o número de observações. Portanto, foi delimitado, para cada iteração, uma amostra aleatória estratificada com reposição de tamanho igual a 3 para cada categoria de pegajosidade.

O valor 3 é escolhido por corresponder ao número de ocorrências da categoria mais rara. Logo, cada árvore de decisão é ajustada para 12 observações, sendo 3 de cada categoria. Esse valor pode ser igual ou menor à categoria de menor ocorrência.

Como apresentado na Figura 12, os erros OOB, são estabilizados antes de 150 árvores agregadas. Em relação à Figura 6, todos os erros são equivalentes ou melhores. Destaca-se a categoria de arroz solto, cujo erro de classificação cai para 0.

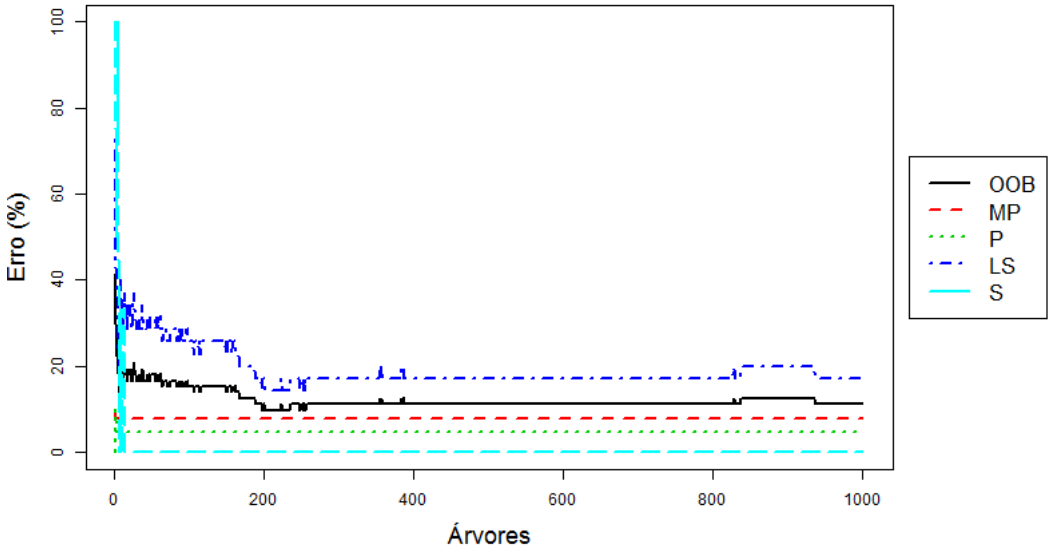


Figura 12 – Erro OOB por Número de Árvores

A Figura 12 apresenta o erro OOB geral e para cada uma das 4 categorias de pegajosidade conforme o número de árvores. O erro estabiliza-se rapidamente, de forma que o acréscimo no número de árvores não influencia na precisão do classificador.

Esse erro está associado ao erro de predição ou generalização, logo, se espera taxas semelhantes de erro para predição de novos dados dada uma mesma distribuição de pegajosidade de arroz.

A Tabela 14 apresenta erros de classificação menores quando comparado com dados desbalanceados. O erro OOB geral foi de 11,11%, isto é, 8 erros em 72 classificações. O padrão de classificação também mudou, foi mais consistente com a característica ordinal dos dados apresentando estimativas sempre em torno da real categoria.

Tabela 14 – Matriz de Confusão para a previsão OOB do modelo de floresta aleatória balanceada

		Classificação Prevista				Erro
		MP	P	LS	S	OBB
Classificação Real	MP	12	1	0	0	7,7%
	P	0	20	1	0	4,8%
	LS	0	3	29	3	17,1%
	S	0	0	0	3	0%

Em seguida, a Figura 13 apresenta a dispersão das observações pelos escores das duas primeiras componentes principais, discriminando as categorias e o tipo de classificação.

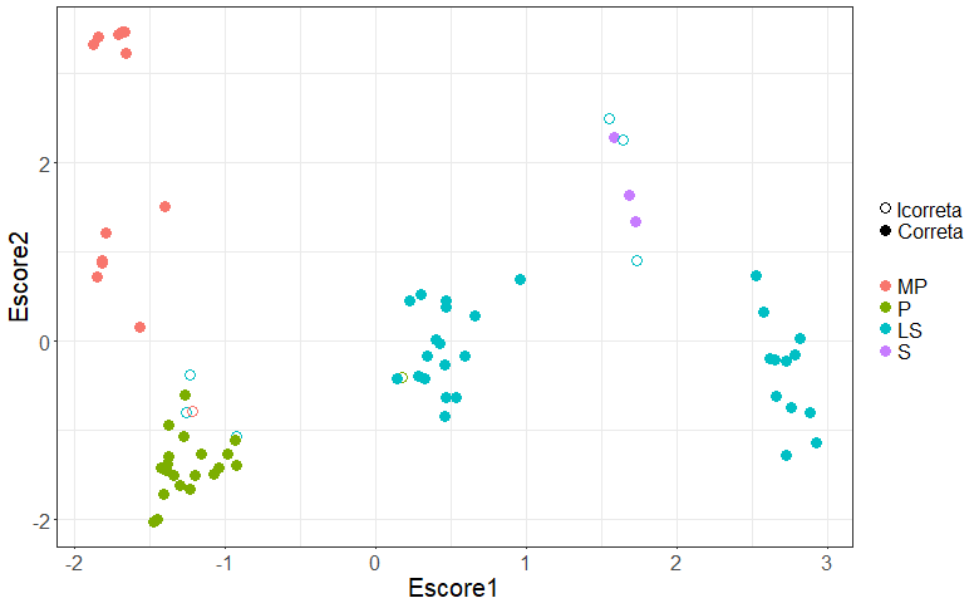


Figura 13 – Dispersão da Classificação Sensorial e Predita com base nos Escores1 e Escores2

As regiões de confundimento de categorias continuam apresentando erros de classificação, mas de forma mais consistente. O modelo identifica melhor os grupos de categorias e atribui erro de classificação aos pontos que naturalmente se destacam.

As probabilidades de classificação podem ser visualizadas por gráficos de calor, conforme a Figura 14.

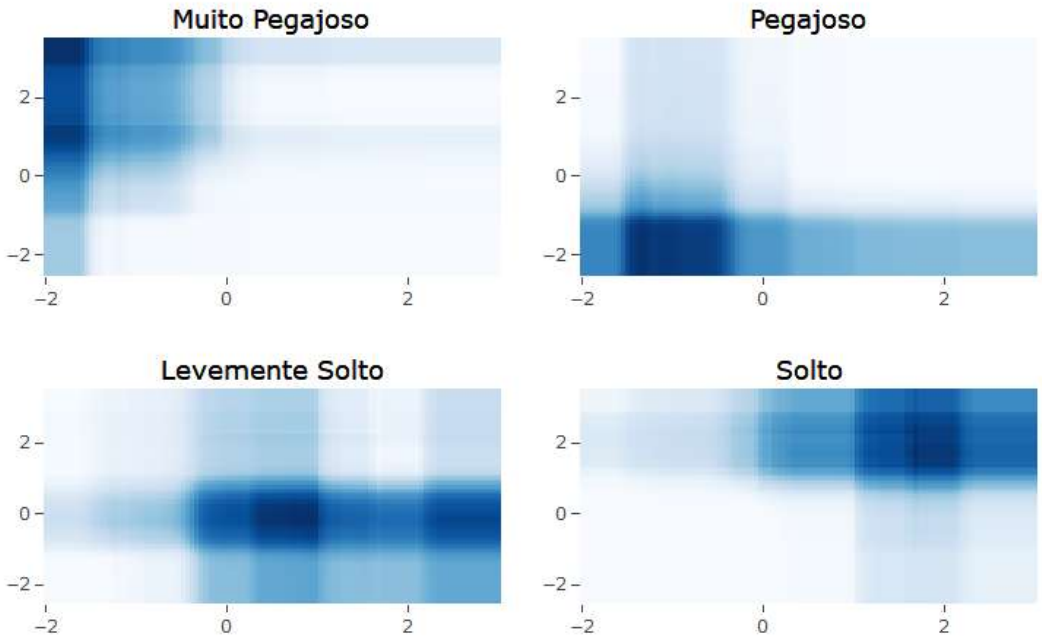


Figura 14 – Gráficos de Calor para Probabilidade de Classificação

As regiões de classificação para cada categoria são mais suaves e apresentam maior distinção para cada gráfico de calor. A categoria solto apresenta uma região de classificação bem definida, diferente da Figura 8. A Figura 14 além de considerar estruturas não lineares de classificação, se aproxima muito à dispersão observada na Figura 13.

Em seguida, Figura 15, são representadas as curvas ROC para cada categoria. Com exceção do nível levemente solto, as curvas mostram alto poder discriminativo, dado a probabilidade estimada via floresta aleatória balanceada.

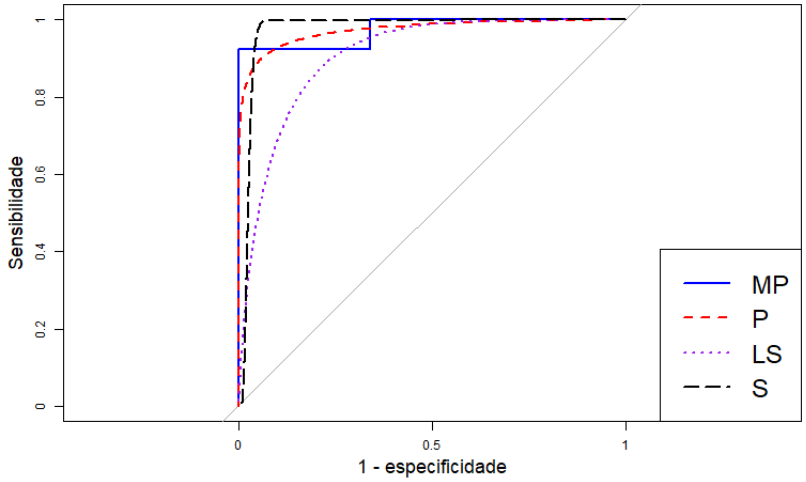


Figura 15 – Curva de classificação ROC da avaliação sensorial de pegajosidade para o ano de 2014, prevista por meio do modelo de floresta aleatória balanceada

A Tabela 15 apresenta as áreas abaixo da curva para cada categoria ajustada pelo modelo para dados desbalanceados.

Tabela 15 – Área sob a curva ROC por categoria de pegajosidade

Modelo	Pegajosidade MP	P	LS	S
Floresta Alea. Balanceada	0,9739	0,9724	0,9068	0,9739

Todos os valores caracterizam excelente poder de discriminação para as categorias de pegajosidade. Indicando que o balanceamento não só favorece categorias de baixa ocorrência, mas ajuda a discriminar as regiões de confundimento.

3.2.2.1 Barreira de Incerteza

Dado o modelo de floresta aleatória balanceada ajustado, são calculados valores de discrepância para cada observação de arroz cozido. A Figura 16 apresenta os valores encontrados e distingue pela cor, se a classificação foi correta, azul, ou incorreta, verde.

Se observa a prevalência da cor verde para os maiores valores de discrepância. Parte das observações que se destacam na Figura 11 também se distinguem para o modelo balanceado.

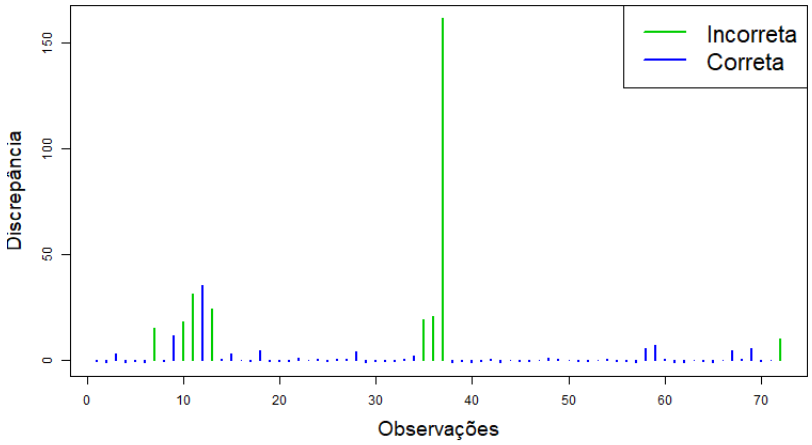


Figura 16 – Valores de discrepância por tipo de classificação

O ponto de corte que maximiza a Equação 2.4, é de $\hat{c} = 12,47$. A matriz de confusão resultante desse corte é apresentada na Tabela 18.

Tabela 16 – Matriz de Confusão para Discrepâncias

		Discrepância		
	$\hat{c} = 12,47$	$D \geq \hat{c}$	$D < \hat{c}$	Total
Classificação	*Não*	7	1	8
Correta	*Sim*	1	63	64

Para auxiliar na avaliação da região de incerteza definida são apresentadas na Tabela 17, métricas de avaliação de barreiras e área sob a curva ROC.

Tabela 17 – Métricas de Regiões de Incerteza

Erro de classif.	Custo	Erro prop.	Classif. correta RI	AUC
1,56%	11,11%	87,50%	12,50%	98,16%

Com o uso da região de incerteza, o erro de classificação pode reduzir para 1,56% a um baixo custo de não classificação. A área sob a curva ROC encontrada foi de 98,16%, indicando alto poder de discriminação da medida de discrepância para classificações corretas e incorretas. O alto valor de erro proporcional mostra que poucos casos falso positivos estão fora da região de incerteza. Por fim, o baixo valor de classificações corretas,

nas regiões de incerteza, indicam que poucas observações verdadeiras positivas foram classificadas como incertas.

Após a exclusão das observações contidas na região de incerteza, a nova matriz de confusão é expressa na Tabela 18.

Tabela 18 – Matriz de Confusão para o modelo de Regressão Logística por Validação Cruzada

		Classificação Prevista				Erro de
		MP	P	LS	S	Classificação
Classificação Real	MP	12	0	0	0	0%
	P	0	18	0	0	0%
	LS	0	1	29	0	3,3%
	S	0	0	0	3	0%

O modelo final, isto é, floresta aleatória balanceada considerando região de incerteza por medidas de discrepância, possui a melhor acurácia. O erro de classificação de quase todas as categorias é zerado, e o nível levemente solto conta com, apenas, 1 erro de classificação.

Esses resultados evidenciam a possibilidade de melhora do modelo de floresta aleatória, tanto por procedimentos de balanceamento, quanto por meio da identificação de regiões de incerteza.

3.3 Resultados Adicionais

O procedimento de estimação e ajuste dos modelos foi reproduzido para os demais subconjuntos de dados de classificação sensorial da pegajosidade do arroz cozido. O número de ocorrências de cada categoria, dado o ano e o tipo de terreno, é apresentado na Tabela 19. Para cada modelo foram registradas a informação de acurácia, via validação cruzada, e métricas de barreiras de incerteza.

Tabela 19 – Classificação sensorial de pegajosidade de arroz cozido

Ano	Tipo de Terreno	EP	MP	P	LS	S
2013	Irrigado	0	31	45	35	6
	Terras Altas	0	12	18	36	6
2014	Irrigado	11	1	19	36	8
	Terras Altas	11	2	21	35	3

Esse método permite verificar o comportamento dos diferentes classificadores em dados distintos, além de permitir uma comparação entre os modelos para cada caso. Os conjuntos são dados em função do ano de coleta e do tipo de terreno.

A Tabela 20 apresenta as taxas de erro de classificação de cada modelo, ajustado a cada conjunto. As categorias de pegajosidade foram agrupadas, em alguns casos, em conformidade com o estudo de Rios (2015).

Tabela 20 – Taxas de erro de classificação via validação cruzada

Ano	Tipo de Terreno	Log	RF	RFB	RBP
2013	Irrigado	**0,50**	0,59	0,57	0,62
	Terras Altas	0,44	**0,43**	0,53	0,62
2014	Irrigado	0,29	**0,20**	0,25	0,25
	Terras Altas	0,14	0,15	**0,12**	0,15
2013 + 2014	Irrigado	**0,47**	0,53	0,53	0,55
	Terras Altas	0,40	**0,36**	0,48	0,47
2013 ⇒ 2014	Irrigado	0,45	0,51	0,47	**0,43**
	Terras Altas	0,74	0,75	0,74	0,75

Nota: Log = Regressão logística politômica ordinal, RF = Floresta aleatória, RFB = Floresta aleatória balanceada, RFP = Floresta aleatória ponderada.

Nota: Negrito para o menor erro para o conjunto.

Os modelos foram ajustados utilizando componentes principais como variáveis preditoras. O número de componentes foi definido de forma a explicar, ao menos, 80% de toda a variância dos dados.

Em negrito estão destacadas os menores valores de erro de previsão para cada subconjunto considerado. As siglas Log, RF, RFB, e RBP correspondem, respectivamente aos modelos: regressão logística politômica ordinal, floresta aleatória, floresta aleatória balanceada e floresta aleatória ponderada.

Foram consideradas 1000 árvores para cada um dos modelos de agregação. O tamanho de estratificação, usado para o balanceamento do RFB, foi definido como o número de ocorrências da categoria de maior raridade. A ponderação do RBP foi calculada como o inverso da taxa de proporções para cada categoria, pois assim, as categorias de menor ocorrência possuem maior peso, enquanto as mais recorrentes possuirão pesos pequenos.

O ano igual a "2013 + 2014" indica que foram utilizados a totalidade dos dados somando os dois anos. A célula "2013 ⇒ 2014" da coluna ano, representa os casos em que o modelo ajustado no ano de 2013 foi usado para a predição dos dados do ano de 2014.

Dos 8 modelos ajustados, a regressão logística obteve desempenho superior apenas em dois casos, arroz de terrenos irrigados para o ano de 2013 e terrenos irrigados quando considerados o total de observações de todos os anos.

Na Tabela 21 são apresentados os valores das métricas de região de incerteza para

a região baseada na discrepância dos diferentes grupos de arroz.

Tabela 21 – Métricas de região de incerteza para o modelo de floresta aleatória

Ano	Tipo de Terreno	Erro de classif.	Custo	Erro prop.	Classif. correta RI	AUC
2013	Irrigado	0,21	0,67	0,89	0,21	0,85
	Terras Altas	0,33	0,28	0,39	0,45	0,65
2014	Irrigado	0,13	0,11	0,40	0,25	0,83
	Terras Altas	0,05	0,12	0,73	0,11	0,88
2013 + 2014	Irrigado	0,23	0,43	0,74	0,12	0,90
	Terras Altas	0,17	0,33	0,70	0,23	0,82
2013 $\Rightarrow$ 2014	Irrigado	0,25	0,73	0,86	0,42	0,71
	Terras Altas	0,75	0,03	0,02	0,50	0,75

As medidas estão diretamente associadas à qualidade do ajuste dos modelos. Verifica-se que ao se considerar o modelo sem as observações, classificadas como incertas, o erro é reduzido para todos os caso. A média de redução de erro foi de 20%. Os custos mais elevados são observados aos casos de menor acurácia, mas esse valor pode ser justificado se conjuntamente se observa alto valor de erro proporcional e baixo valor para as classificações corretas contidas na região de incerteza. Por fim, os altos valores de áreas abaixo da curva corroboram com a intuição do erro de classificação ser bem relacionado ao valor de discrepância, via proximidades da floresta aleatória.

O resultado da barreira de incerteza se mostra consistente, inclusive ao cenário de predição para novos dados. O procedimento de barreiras via discrepâncias é eficaz na redução do erro de classificação, por discriminar bem as classificações corretas das incorretas.

4 Discussão e Conclusões

O trabalho explorou o uso das florestas aleatórias para o caso de classificação com aplicação à análise sensorial da pegajosidade de amostras de arroz cozido. Três tópicos guiaram a aplicação da técnica: busca por modelos bem ajustados, apresentação gráfica para classificadores complexos e determinação de barreiras consistentes para regiões de incerteza.

Os modelos de classificação gerados pelas florestas aleatórias se mostraram competitivos, por vezes superiores, aos modelos gerados via regressão logística politômica ordinal. Mesmo sem considerar métodos para tratar o desbalanceamento, as florestas aleatórias já apresentavam erros mais balanceados entre as classes, diferentes da regressão logística cujos melhores ajustes foram observados para as classes mais frequentes.

Ao considerar a floresta aleatória balanceada para o arroz, de terras altas de 2014, se observou ganhos significativos para a acurácia de todas as categorias. Para os demais conjuntos de dados a floresta aleatória apresentou maior acurácia na maioria dos casos. Isto pode indicar que métodos de balanceamento necessitam ser aplicados a casos específicos.

No tocante da visualização das probabilidades de classificação geradas pelo modelo de floresta aleatória, evidenciou-se algumas peculiaridades. A forma da função de probabilidade não é contínua e sim dada em saltos de probabilidade, isso se deve a estimação ser dada de maneira empírica por meio de médias de proporções. As regiões de probabilidade são apresentadas em formatos retangulares, visto que cada árvore de decisão é baseada no seccionamento simples e recursivo do espaço formado pelas variáveis explicativas.

A medida de discrepância considerada parece sempre estar relacionada à discriminação de observações correta e incorretamente classificadas. O critério para identificação de incerteza é muito objetivo e independe de quais categorias o modelo possui. Verificou-se que o desempenho da barreira está muito atrelado à qualidade do ajuste do modelo.

A princípio, a análise prevê respostas sensoriais numa escala de Likert de 7 níveis. Mesmo quando considerados outros conjuntos de dados disponíveis, isto é, dados para os anos de 2013 e 2014, para terrenos irrigados e terras altas, não se observa a ocorrência de ao menos duas categorias. A maioria dos modelos de classificação é incapaz de prever alguma probabilidade de ocorrência para classes não observadas. A modelagem Bayesiana é exceção à essa limitação, mas Oliveira (2015) mostrou que não há ganho na precisão quando considerado esse paradigma. Portanto, independente do modelo utilizado, a predição para as classes não observadas está comprometida para os dados coletados.

Apesar da insensibilidade (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) do modelo

de floresta aleatória para variações no número de variáveis a ser consideradas em cada partição, profundidade das árvores de decisão e tamanho da amostra *bootstrap*, Biau e Scornet (2016) argumentam que esses valores são parâmetros e por isso também deveriam ser calibrados para a melhor modelagem. Sendo assim, um exercício possível é testar diferentes valores desses parâmetros iniciais e seu impacto no modelo final. O tratamento para dados desbalanceados sofre de crítica semelhante, e nesse caso, de fato é mais sensível aos valores de ponderação ou custo escolhidos.

Florestas aleatórias utilizam o critério de classificação das árvores de decisão. Nesse sentido, não se consideram padrões ordinais na resposta categórica. Hu et al. (2011) propõem uma medida de informação ordinal para construção das árvores de decisão, cujos resultados são favoráveis à melhor classificação. Portanto, a extensão para uma floresta aleatória ordinal deve proporcionar um modelo mais recomendado.

O critério estudado para identificação de região de incerteza apresenta bons resultados às aplicações consideradas. Um próximo passo é reproduzir os demais tipos de barreiras, propostos por Rocha (2017), para comparações de desempenho. Estudos em dados simulados permitiriam a avaliação desse método para diferentes cenários controlados, o que corroboraria com sua eficácia.

As técnicas de classificação utilizadas também são apresentadas no cenário de aprendizado de máquina, que é caracterizada, principalmente, por sua autonomia e aplicações que dispensam pré-tratamentos aos dados, por exemplo, de eliminação de multicolinearidade e ruído. Outras duas técnicas de classificação se destacam e talvez apresentem soluções interessantes aos dados, são elas: máquinas de suporte vetorial (*Support Vector Machine - SVM*, em inglês) e redes neurais (*Neural Network*, em inglês).

Referências

AGRESTI, A. *Categorical Data Analysis*. New Jersey: Wiley, 2003. (Wiley Series in Probability and Statistics). ISBN 9780471458760. Citado na página 15.

BIAU, G.; SCORNET, E. A random forest guided tour. *TEST*, v. 25, n. 2, p. 197–227, Jun 2016. ISSN 1863-8260. Disponível em: <https://doi.org/10.1007/s11749-016-0481-7>. Citado na página 54.

BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, Aug 1996. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1018054314350>. Citado na página 19.

BREIMAN, L. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, The Institute of Mathematical Statistics, v. 24, n. 6, p. 2350–2383, 12 1996. Disponível em: <https://doi.org/10.1214/aos/1032181158>. Citado 3 vezes nas páginas 18, 19 e 21.

BREIMAN, L. *Out-of-bag estimation*. 1996. Citado na página 19.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. Citado 3 vezes nas páginas 21, 22 e 33.

BREIMAN, L. *Setting up, using, and understanding random forests V4.0*. 2003. Disponível em: <https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf>. Citado na página 23.

BREIMAN, L. et al. *Classification and Regression Trees*. Belmont, California, U.S.A.: Wadsworth Publishing Company, 1984. (Statistics/Probability Series). Citado 2 vezes nas páginas 15 e 18.

CHEN, C.; LIAW, A.; BREIMAN, L. *Using random forest to learn imbalanced data*. UC Berkeley, 2004. Citado na página 24.

EFRON, B.; TIBSHIRANI, R. J. *An Introduction to the Bootstrap*. Boca Raton, Florida, USA: Chapman & Hall/CRC, 1993. (Monographs on Statistics and Applied Probability, 57). Citado na página 19.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, Blackwell Publishing Ltd, v. 7, n. 2, p. 179–188, 1936. ISSN 2050-1439. Disponível em: <http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x>. Citado 3 vezes nas páginas 16, 25 e 31.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference and prediction*. 2. ed. New York: Springer, 2009. Citado 2 vezes nas páginas 22 e 53.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. Canada: John Wiley and Sons, 2000. ISBN 0471356328, 9780471356325. Citado 2 vezes nas páginas 14 e 15.

HU, Q. et al. Rank entropy-based decision trees for monotonic classification. *IEEE*, 2011. Citado na página 54.

JAMES, G. et al. *An Introduction to Statistical Learning: with Applications in R*. New York: Springer, 2014. (Springer Texts in Statistics). ISBN 9781461471370. Citado na página 16.

JOHNSON, R.; WICHERN, D. *Applied Multivariate Statistical Analysis*. New Jersey: Pearson Prentice Hall, 2007. Citado 3 vezes nas páginas 13, 14 e 27.

LOH, W. Y.; SHIH, Y. S. Split selection methods for classification trees. *Statistica Sinica*, Institute of Statistical Science, Academia Sinica, v. 7, n. 4, p. 815–840, 1997. ISSN 10170405, 19968507. Citado na página 18.

MARONNA, R.; MARTIN, D.; YOHAI, V. *Robust Statistics: Theory and Methods*. Chichester: Wiley, 2006. (Wiley Series in Probability and Statistics). ISBN 9780470010921. Citado na página 23.

OLIVEIRA, G. S. Modelos de regressão com resposta ordinal para avaliação de textura de arroz. *Trabalho de conclusão de curso, Departamento de Estatística, UnB*, 2015. Citado 4 vezes nas páginas 12, 35, 36 e 53.

RIOS Érica S. Modelos estatísticos para avaliação da qualidade culinária de arroz: Textura e propriedades viscoamilográficas. *Trabalho de conclusão de curso, Departamento de Estatística, UnB*, 2015. Citado 7 vezes nas páginas 12, 25, 28, 35, 36, 40 e 50.

ROCHA, L. T. Estudo de regiões de incerteza na avaliação e ajuste de escalas de classificação sensorial de arroz. *Trabalho de conclusão de curso, Departamento de Estatística, UnB*, 2017. Citado 8 vezes nas páginas 11, 12, 29, 31, 35, 37, 40 e 54.

TIBSHIRANI, R. *Bias, Variance and Prediction Error for Classification Rules*. 1996. Citado na página 19.

WOLPERT, D. H.; MACREADY, W. G. An efficient method to estimate bagging's generalization error. *Machine Learning*, v. 35, n. 1, p. 41–55, Apr 1999. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1007519102914>. Citado na página 19.

ZHOU, Z. *Ensemble Methods: Foundations and Algorithms*. Boca Raton, Florida, USA: Chapman and Hall/CRC, 2012. Citado na página 18.

Apêndices

APÊNDICE A – Visualização de Probabilidade dado o Número de Árvores Combinadas

Para facilitar o entendimento do mecanismo de classificação utilizado pela floresta aleatória são apresentados gráficos de calor para a probabilidade de classificação dado o número de árvores consideradas.

Foram considerados dados simulados com alto confundimento e não linearmente separáveis. A Figura 17 apresenta a dispersão das quatro categorias de dados, para as duas variáveis explicativas.

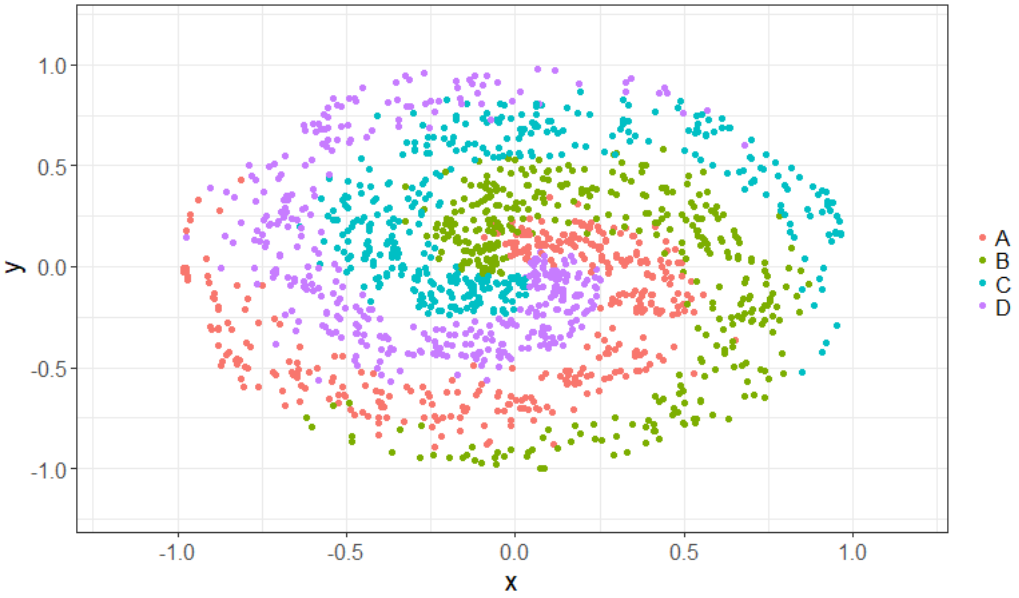


Figura 17 – Dispersão de dados em espiral

São 800 pontos que estão balanceados para as 4 categorias, ou seja, 200 observações para cada nível. A probabilidade de classificação para floresta aleatória quando considerada apenas uma árvore é apresentada na Figura 18. Com 1 árvore, o modelo de floresta aleatória é equivalente ao ajuste de árvore de decisão sem podas. Apesar da estrutura não linear dos dados, a árvore de classificação capta o comportamento espiral já na primeira árvore. Por conta do sobreajuste as regiões de cor azul escuro correspondem à probabilidade igual a 1 de classificação, as demais regiões possuem probabilidade nula.

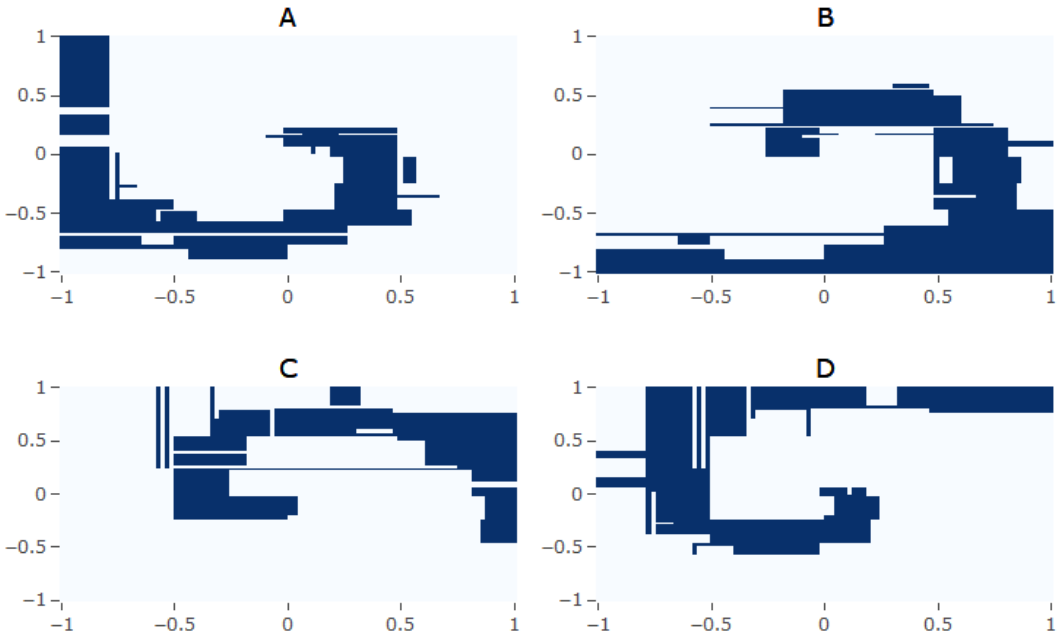


Figura 18 – Gráficos de Calor para Probabilidade de Classificação - 1 árvore

Em seguida, Figura 19, é apresentada a probabilidade de classificação para duas árvores agregadas. Esse gráfico de calor já apresenta probabilidades intermediárias de classificação dada pela combinação de duas árvores distintas.

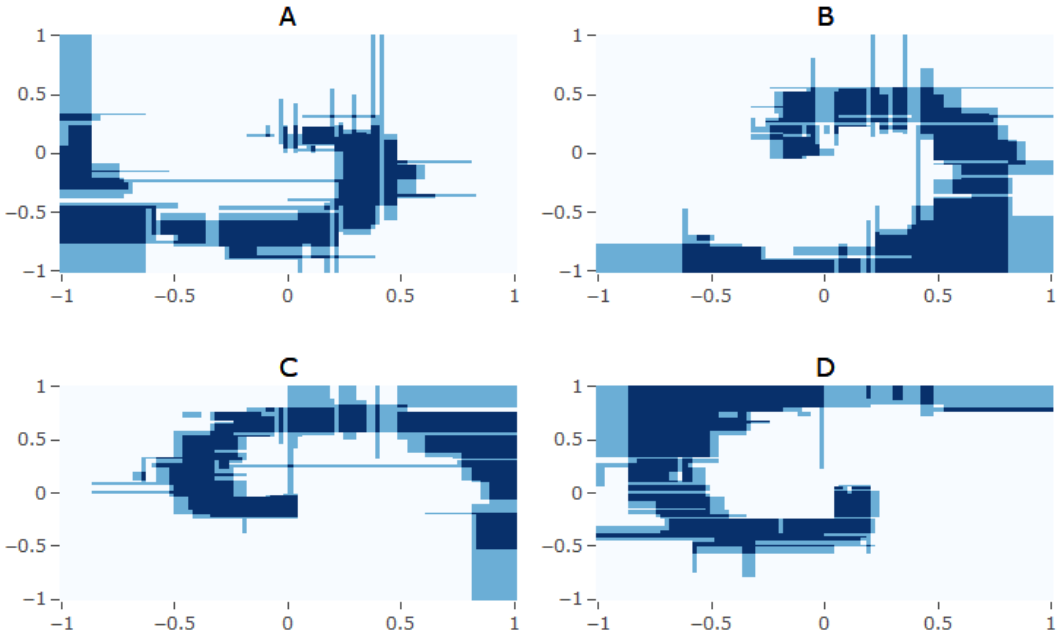


Figura 19 – Gráficos de Calor para Probabilidade de Classificação - 2 árvores

A Figura 20, representa os gráficos de calor para uma floresta aleatória de 4 árvores

de classificação. O ganho na suavização é bem evidente fazendo com que a predição se assemelhe muito à dispersão observada na Figura 17

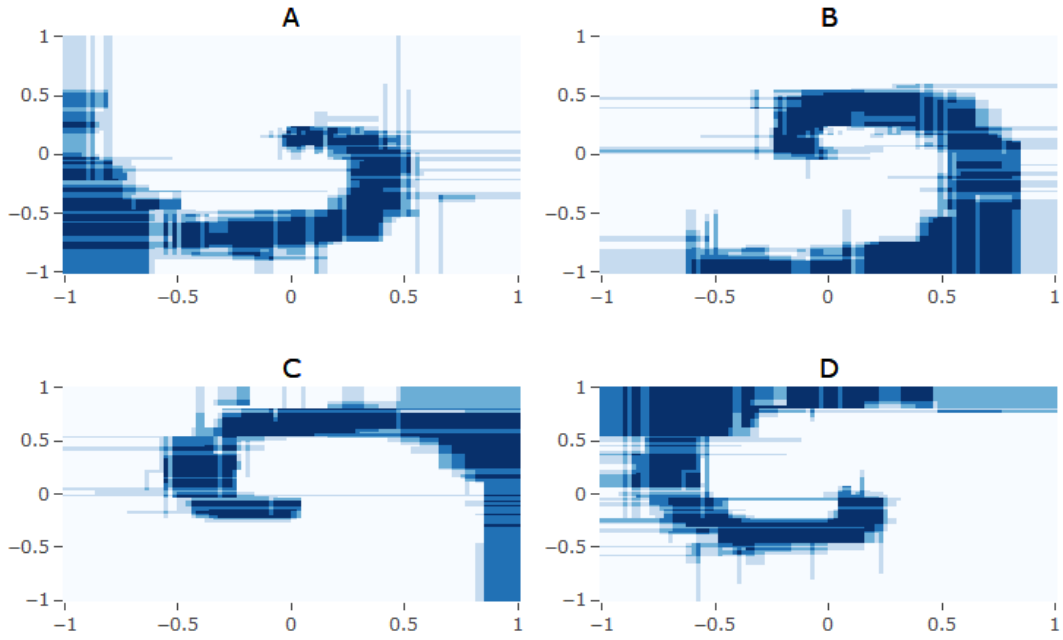


Figura 20 – Gráficos de Calor para Probabilidade de Classificação - 4 árvores

Por fim, a Figura 21 expressa as probabilidades de classificação para o modelo de floresta aleatória com 1000 árvores.

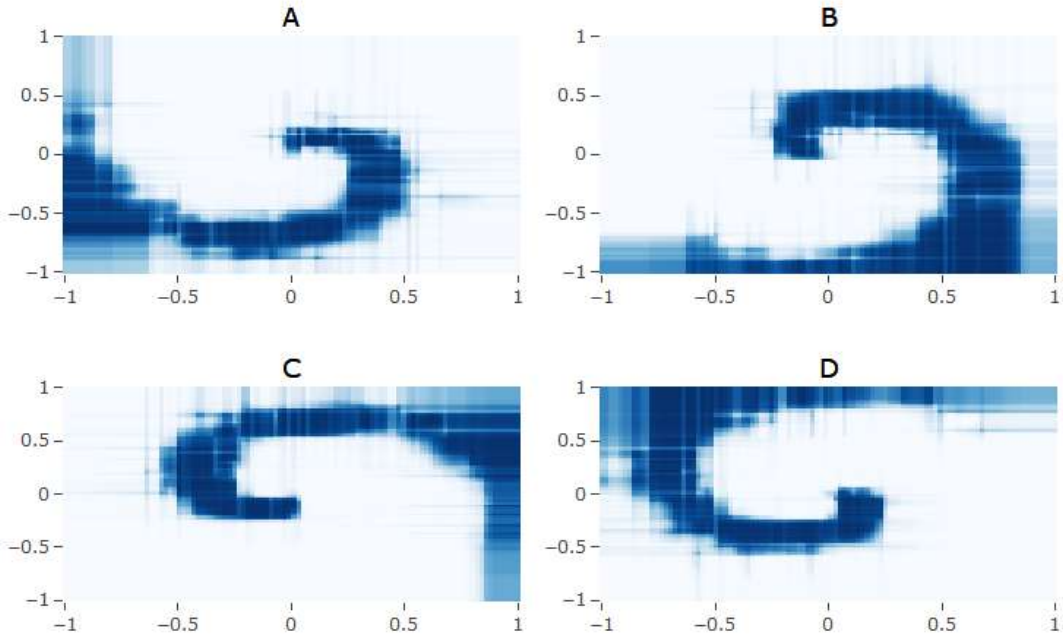


Figura 21 – Gráficos de Calor para Probabilidade de Classificação - 1000 árvores

Apesar da complexidade, o modelo não sobreajusta, o que permite boas estimativas de generalização.

APÊNDICE B – Códigos R

B.1 Construção de probabilidades de floresta aleatória

```
install.packages('ggplot2')
library(ggplot2)

N <- 200     # Número de pontos por categoria
K <- 4       # Número de categorias
X <- data.frame()
y <- data.frame()

# Simulação de dados espirais
set.seed(321)
for (j in (1:K)){
  r <- seq(0.05,1,length.out = N)
  t <- seq((j-1)*4.7,j*4.7, length.out = N) + rnorm(N, sd = 0.4)
  Xtemp <- data.frame(x =r*sin(t) , y = r*cos(t))
  ytemp <- data.frame(matrix(j, N, 1))
  X <- rbind(X, Xtemp)
  y <- rbind(y, ytemp)
}

data <- cbind(X,y)
colnames(data) <- c(colnames(X), 'label')

x_min <- min(X[,1]) -0.2; x_max <- max(X[,1])+0.2
y_min <- min(X[,2]) -0.2; y_max <- max(X[,2])+0.2

data$label <- c('A', 'B', 'C', 'D')[data$label] # Categorias


# Gráfico de dispersão
ggplot(data) + geom_point(aes(x=x, y=y, color = as.character(label)), size
   = 2) +
  theme_bw(base_size = 15) + xlim(x_min, x_max) + ylim(y_min, y_max) +
  guides(colour=guide_legend(title=NULL)) +
  theme(text = element_text(size=20))


install.packages('randomForest')
library(randomForest)

# Modelo com 1 árvore
```

```
rf1 <- randomForest(factor(label) ~ x + y, data, ntree = 1)

# Modelo com 2 árvores
rf2 <- randomForest(factor(label) ~ x + y, data, ntree = 2)

# Modelo com 4 árvores
rf4 <- randomForest(factor(label) ~ x + y, data, ntree = 4)

# Modelo com 1000 árvores
rf1000 <- randomForest(factor(label) ~ x + y, data, ntree = 1000)

# Função para gráficos de calor
install.packages('plotly')
heatmaps <-  function(rf){
x <- seq.int(-1, 1, length.out =  100)
w <- x
x <- expand.grid(x = x, y = x)

z <- array((predict(rf, x, type="prob")), dim = c(100, 100, 4))

# Parâmetros gráficos
f <- list( size = 18, color = "black")

x <- list( title = "x Axis", titlefont = list(size = 16) )
y <- list( title = "y Axis", titlefont = list(size = 16) )

a1 <- list( text = "A", font = f, xref = "paper",
yref = "paper", yanchor = "bottom", xanchor = "center",
align = "center", x = 0.5, y = 1, showarrow = FALSE )

a2 <- list( text = "B", font = f, xref = "paper",
yref = "paper", yanchor = "bottom", xanchor = "center",
align = "center", x = 0.5, y = 1, showarrow = FALSE )

a3 <- list(  text = "C", font = f, xref = "paper",
yref = "paper", yanchor = "bottom", xanchor = "center",
align = "center", x = 0.5, y = 1, showarrow = FALSE )

a4 <- list( text = "D", font = f, xref = "paper",
yref = "paper", yanchor = "bottom", xanchor = "center",
align = "center", x = 0.5, y = 1, showarrow = FALSE )

library(plotly)

pr <- list()
pr[[1]] <- plot_ly(showscale = F, colors="Blues",
hoverlabel=list(namelength='-1')) %>%
```

```
add_heatmap(x = w, y = w, z = ~t(z[,,1]))%>%
layout(annotations = a1)

pr[[2]] <- plot_ly(showscale = F, colors="Blues", name = ~'Pegajoso',
hoverlabel=list(namelength='-1')) %>%
add_heatmap(x = w, y = w, z = ~t(z[,,2]))%>%
layout(annotations = a2)

pr[[3]] <- plot_ly(showscale = F, colors="Blues", name = ~'Levemente<br>
    Solto',
hoverlabel=list(namelength='-1')) %>%
add_heatmap(x = w, y = w, z = ~t(z[,,3]))%>%
layout(annotations = a3)

pr[[4]] <- plot_ly(showscale = F, colors="Blues", name = ~'Solto',
hoverlabel=list(namelength='-1')) %>%
add_heatmap(x = w, y = w, z = ~t(z[,,4]))%>%
layout(annotations = a4)

return(
  do.call(subplot, c(pr, list(nrows = 2, titleX = F,
        titleY = F, margin = c(0.04, 0.04, .12, .08))))
  )
}

# Gráficos de calor
heatmaps(rf1)
heatmaps(rf2)
heatmaps(rf4)
heatmaps(rf1000)
```

B.2 Árvore de Decisão - Iris

```
install.packages("tree")
library(tree)

# Ajuste de árvore de decisão
iris.tree <- tree(Species ~ Petal.Length + Petal.Width, data=iris)

iris$Comprimento <- iris$Petal.Length
iris$Largura     <- iris$Petal.Width

# Representação Gráfica
par(mfrow=c(1,2))
plot(iris.tree)
text(iris.tree, cex=.75)
```

```
plot(iris[c("Petal.Length", "Petal.Width")], type="n")
text(iris[c("Petal.Length", "Petal.Width")], c("s", "c", "v")[iris[, 5]],
col=c("red", "blue", "green")[iris[, 5]])
partition.tree(iris.tree, add = T, cex = .8)
```

B.3 Gráfico de Calor - Iris

```
install.packages('randonForest')
library(randomForest)

# Ajuste de Floresta Aleatória
set.seed(321)      # Necessário para fixar o resultado
rf.iris <- randomForest(Species ~ Petal.Length + Petal.Width, data=iris,
ntree = 1000)

summary(iris[,c("Petal.Length", "Petal.Width")])     # Checa a Amplitude

# Pontos do gráfico
x1 <- seq.int(1, 6.9, length.out =  100)
x2 <- seq.int(0.01, 2.5, length.out =  100)

x <- expand.grid(Petal.Length = x1, Petal.Width = x2)

w <- array((predict(rf.iris, x, type="prob")), dim = c(100, 100, 3))

# Representação gráfica
install.packages('plotly')
library(plotly)

f <- list( size = 18, color = "black")

b1 <- list(text = "setosa", font = f, xref = "paper",
       yref = "paper", yanchor = "bottom", xanchor = "center",
       align = "center", x = 0.5, y = 1, showarrow = FALSE )

b2 <- list( text = "versicolor", font = f, xref = "paper",
       yref = "paper", yanchor = "bottom", xanchor = "center",
       align = "center", x = 0.5, y = 1, showarrow = FALSE )

b3 <- list( text = "virginica", font = f, xref = "paper",
       yref = "paper", yanchor = "bottom", xanchor = "center",
       align = "center", x = 0.5, y = 1, showarrow = FALSE )

ps <- list() # Lista que armazena cada gráfico de calor

ps[[1]] <- plot_ly(showscale = F, colors="Blues",
    hoverlabel=list(namelength='-1')) %>%
```

```
    add_heatmap(x = ~x1, y = ~x2, z = ~t(w[,,1])) %>%
    layout(annotations = b1)

ps[[2]] <- plot_ly(showscale = F, colors="Blues",
    hoverlabel=list(namelength='-1')) %>%
    add_heatmap(x = x1, y = x2, z = ~t(w[,,2])) %>%
    layout(annotations = b2)

ps[[3]] <- plot_ly(showscale = F, colors="Blues",
    hoverlabel=list(namelength='-1')) %>%
    add_heatmap(x = x1, y = x2, z = ~t(w[,,3])) %>%
    layout(annotations = b3)

# Gráficos na mesma figura
do.call(subplot, c(ps, list(titleX = F, titleY = F,
          margin = c(0.04, 0.04, .12, .08))))
```

B.4 Barreira de Incerteza - Iris

```
install.packages('randomForest')
library(randomForest)

set.seed(321)
rf.iris <- randomForest(Species ~ Petal.Length + Petal.Width, data=iris,
             proximity=T, ntree = 1000)

clas.pred      <- predict(rf.iris) == iris$Species
out            <- outlier(rf.iris)
out[is.na(out)] <- 0      # Valores NA's surgem quando MAD = 0

# Gráfico de valores de discrepância
plot(out, type="h", col = palette()[3+clas.pred], lwd=2,
xlab = "Observações", ylab = "Discrepância", cex.lab=1.6)
legend("topleft", cex=1.6,
legend=c("Incorreta", "Correta"), col=palette()[3:4], lwd=2)

# Calculo do melhor ponto de corte
erro <- vector()
for(i in 1:length(clas.pred)) {
  c       <- sort(out)[i]
  conf    <- table(clas.pred, ifelse(out < c, T, F))
  erro[i] <- sum(diag(conf)) - sum(conf)
}

(t <- (which(erro==max(erro)))) # Observação que determina o corte
(A <- sort(out)[t])             # Ponto de corte ótimo
```

```
# Matriz de confusão do ponto ótimo de corte
(con <- table(clas.pred, ifelse(out < A, T, F)))

# Métricas de barreira de incerteza
con[1,2]/colSums(con)[2]  # Erro apar.
colSums(con)[1]/sum(con)  # Custo
con[1,1]/rowSums(con)[1]  # Erro prop.
con[2,1]/colSums(con)[1]  # Clas. Cor. RI

# Curva ROC
install.packages('pROC')
pROC::roc(clas.pred, out, smooth=T, levels=c(F,T))
```