Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Estudo de Regiões de Incerteza na Avaliação e Ajuste de Escalas de Classificação Sensorial de Arroz

Laura Teixeira da Rocha

Brasília 2016

Estudo de Regiões de Incerteza na Avaliação e Ajuste de Escalas de Classificação Sensorial de Arroz

Laura Teixeira da Rocha

Brasília 2016

Relatório apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos para obtenção de título de Bacharel em Estatística.

Orientador: Prof.° George Freitas von Borries

Agradecimentos

À minha mãe Sandra, que sempre foi e sempre será minha alegria, exemplo de positividade e fonte de inspiração. Ao meu pai Adson, que sempre me mostrou o caminho da paciência, honestidade e dedicação. À minha irmã, Cecília, pelo seu companheirismo e por sempre ter sido meu exemplo de sucesso.

Ao Prof. George pela excelente orientação e paciência dedicados durante esse período essencial da minha graduação, tornando o fim desse ciclo uma experiência prazerosa e de imensa agregação. Aos profissionais da Embrapa, pelo empenho e envolvimento em seu trabalho e por tornarem a realização deste estudo possível. Aos docentes e funcionários da Universidade de Brasília, em especial ao Departamento de Estatística, por acreditarem e fazerem parte do processo da educação do país e pela atenção e carinho dedicados aos alunos todos os dias.

Aos meus companheiros de vida, que me conhecem melhor do que ninguém: Bruno Vilas Boas, Eduarda Leão, Marina Helena e Marina Macedo. Vocês são a minha definição de amizade e cúmplicidade. Diversos obstáculos apareceram durante essa trajetória e somente puderam ser superados por vocês estarem presentes em minha vida. E mais importante, obrigada pelas risadas incansáveis e por me proporcionarem lembranças inesquecíveis.

Ao meu amigo Lucas Rodrigues, que teve influência fundamental para que este trabalho pudesse ser concluído. Obrigada pela paciência, tempo e compreensão e, principalmente, pelas risadas e companheirismo de todo dia. Às minhas amigas Gabriela Turquetti, Laís Sakkis, Mariana de Abreu e Thayanne Sales, pelos anos de amizade e por sempre se fazerem presentes em momentos difíceis.

À ESTAT Consultoria, por me dar a oportunidade de me capacitar profissionalmente e sem dúvidas me proporcionar os melhores anos da minha graduação. Por fim, gostaria de agradecer todos àqueles que passaram por minha vida e de alguma maneira deixaram em mim um pouquinho de si e me fizeram quem eu sou hoje.

Resumo

O estudo da análise da textura de arroz pode ser realizado por meio da avaliação sensorial desse grão. Esse método exige que profissionais sejam bem treinados e possui limitações, como o alto custo e a necessidade de alta quantidade de tempo para sua realização. No trabalho, utilizando-se um modelo estatístico criado para a automatização desse processo, é realizada a previsão das classificações sensoriais do arroz de maneira rápida e com custos menores. Entretanto, após a previsão do modelo, é possível observar regiões de incerteza onde ocorrem erros de classificação. O objetivo deste trabalho é fazer o estudo dessas regiões de incerteza de classificação do arroz segundo a classificação sensorial e a classificação conduzida pelo modelo estatístico. Assim, são utilizadas técnicas de classificação de dados e técnicas não encontradas na literatura até então, para a definição das regiões de incerteza, e assim, fazer a redução dos erros de classificação. Os resultados mostraram que para critérios que identificam um maior número de erros de classificação, possuem um custo mais elevado, em que muitas amostras classificadas corretamente são observadas nessas regiões. As regiões de incerteza que identificam uma quantidade menor de erros de classificação são mais vantajosas, uma vez que menos amostras são classificadas como incerteza, e assim, possuem um erro de classificação menor do que o observado originalmente.

Palavras-chave: Textura de arroz, avaliação sensorial, classificação de dados, regiões de incerteza, erro de classificação.

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução e justificativa

O estudo da qualidade do arroz é de extremo interesse dos fabricantes e consumidores desse grão. A análise da textura de arroz é uma das principais características a serem estudadas, pois ela influencia a comercialização desse produto. A avaliação da textura do grão de arroz pode ser feita de três maneiras: por meio da avaliação sensorial da textura de arroz, via medidas instrumentais da textura de arroz e a partir de medidas instrumentais de viscosidade de arroz. Por isso, estudos foram conduzidos por Rios [1] e Oliveira [2], visando criar modelos que pudessem prever a qualidade do arroz com maior eficiência do que os métodos utilizados até então. Esses estudos representaram avanços na otimização de tempo e custo para a obtenção de tais resultados. Ambos os trabalhos utilizam as técnicas de componentes principais e regressão logística, sendo que o primeiro utiliza a visão estatística clássica, e o segundo, a estatística bayesiana. Entretanto, os resultados obtidos mostraram que ainda há estudos que podem ser feitos, para que o modelo criado atinja o seu objetivo de maneira ainda mais eficiente. O objetivo principal deste trabalho é fazer o estudo de regiões de incerteza geradas por erros de classificação, categorias de probabilidades estimadas semelhantes, entre outros. Para este estudo são utilizados modelos propostos por Rios.

A Empresa Brasileira de Pesquisa Agroprecuária (Embrapa) é uma empresa voltada ao desenvolvimento de tecnologias, que visam ao constante progresso na agricultura e pecuária brasileira. Conforme descrito anteriormente, o arroz é um grão essencial e necessário no dia-a-dia do brasileiro e a avaliação sensorial de sua textura é realizada em processo desgastante e de alto custo, uma vez que os profissionais que a executam devem ser bem treinados e a coleta de amostras requer muito tempo. Uma vez que o processo depende da sensibilidade do tato do ser humano, uma quantidade restrita de amostras pode ser analisada por dia. Após diversas análises, o tato do avaliador se acomoda à textura do arroz, podendo haver a ocorrência maior de erros na definição das categorias. Dessa forma, Rios propôs um modelo que viabiliza a substituição da análise sensorial por meio do estabelecimento de uma relação entre medidas sensoriais e medidas de perfil viscoamilográfico do arroz.

O processo de avaliação sensorial pelos profissionais é realizado da seguinte maneira: o arroz é cozido por determinado tempo e temperatura; as amostras são apresentadas aos avaliadores em cabines separadas e estes, usando o tato, determinam e registram

a categoria à qual cada amostra pertence em uma ficha designada para tal. A Figura 1.1 mostra, respectivamente, o processo descrito.

Figura 1.1: Processo de Classificação Sensorial do arroz conduzido na Embrapa.

As variáveis pegajosidade e dureza são as variáveis que permitem as análises sensoriais do arroz. Por convenção inicial, no estudo realizado pela Embrapa, a cada uma dessas variáveis pode ser atribuído um valor em uma escala de 1 a 7. Para o caso da avaliação da pegajosidade do arroz, a classificação ocorre de acordo com o quadro a seguir. Uma escala análoga é também definida para a dureza.

Tabela 1.1: Escalas de avaliação da pegajosidade de arroz.

Escala	Classificação
1	Extremamente solto
2	Muito solto
3	Solto
4	Ligeiramente solto
5	Pegajoso
6	Muito pegajoso
7	Extremamente pegajoso

Vale ressaltar que a Figura 1.1 mostra a ficha de registro da coleta de dados que a Embrapa está conduzindo atualmente. A Figura 1.2 mostra exemplos de amostras de arroz determinadas como Solto e Muito Pegajoso.

Figura 1.2: Intrumentos utilizados para medição da variável TG (temperatura de gelatinização).

Durante o estudo de Rios, notou-se a ocorrência de problemas devidos às escalas definidas e utilizadas na Tabela 1.1. Por isso, a redução do número de escalas foi necessária para a criação do modelo. Após a criação deste, foi possível observar a presença de incertezas na classificação das amostras para duas situações: (i) classificação do arroz com base na avaliação sensorial e (ii) com base no modelo utilizado para a previsão destes. Sendo assim, o objetivo deste trabalho é fazer o estudo dessas regiões de incerteza.

Além disso, uma tentativa de redefinição das escalas está sendo realizada pela Embrapa. A colheita de um novo banco de dados está sendo feita com base nas novas escalas, buscando suprir as limitações surgidas em decorrência das categorias originalmente utilizadas. Durante este trabalho foi realizado um estudo que permitiu o entendimento completo do modelo e da problematização, content de modo que o objetivo proposto pudesse ser atingido de maneira eficiente.

2 Revisão Metodológica

Dada a importância e a grande utilização do arroz no Brasil, a Embrapa criou um projeto específico voltado para o estudo desse grão, chamado projeto QualiArroz. Os dados utilizados por Rios e que serão utilizados neste trabalho foram fornecidos pela Embrapa Arroz e Feijão. Os dados foram colhidos ao longo dos anos de 2013 e 2014.

Conforme citado anteriormente, a análise de textura de arroz pode ser realizada por meio de medidas sensoriais de textura, medidas instrumentais de textura e medidas de perfil viscoamilográfico. O banco de dados possui 18 variáveis, sendo nove qualitativas e nove quantitativas. Entre estas, duas das variáveis qualitativas representam medidas sensoriais, a pegajosidade e a dureza, medidas em uma escala de 1 a 7. As medidas instrumentais são representadas também pela dureza e pela pegajosidade. Nesse caso, as duas variáveis são medidas quantitativamente. Sete das variáveis quantitativas do banco de dados representam medidas de perfil viscoamilográfico, fornecendo informações a respeito da viscosidade e teor de amilose aparente deste grão. Por fim, existem variáveis auxiliares que representam como se deu o plantio de arroz. Entre essas, a variável tipo de terreno é de grande importância para este estudo. Os tipos de terrenos dividem-se em terrenos irrigados e terras altas.

Na análise de Rios, pôde-se perceber que certas categorias apresentavam baixa frequência de observação de dados. Sendo assim, categorias com poucas observacões foram retiradas, e as sete categorias apresentadas na Tabela 1.1 foram reduzidas a apenas quatro:

Tabela 2.1: Escalas de avaliação da pegajosidade de arroz após redução.

Escala	Classificação
1	Solto
2	Ligeiramente solto
3	Pegajoso
4	Muito pegajoso

Além disso, regiões de incerteza para a classificação das amostras de arroz são comumente geradas, em razão de dúvida do profissional ou pela igual probabilidade de pertencer a distintas categorias, calculada no modelo de Rios. Neste, a definição das regiões de incerteza será realizada, no intuito de aumentar o acerto na categorização das amostras de arroz e, consequentemente, reduzir o erro de classificacão. As regiões de

incerteza serão explicadas detalhadamente neste trabalho.

A análise de componentes principais busca explicar a matriz de variância-covariância de determinadas variáveis de um conjunto de dados, por meio de combinações lineares dessas variáveis. O objetivo geral dessa análise busca a redução do banco de dados e a interpretação deste, com base no estudo que se deseja realizar. Após a utilização de componentes principais, e feita a devida redução das variáveis explicativas, é aplicada a regressão logística para a criação de modelos.

A análise de regressão é um método estatístico que utiliza da relação entre duas ou mais variáveis, de modo que a variável resposta possa ser predita pelas outras [8]. As variáveis reposta nesse estudo são as medidas sensoriais de textura, pegajosidade e dureza. Como essas variáveis são colocadas em categorias, se faz necessário o uso da regressão logística, uma vez que a regressão linear exige que os dados referentes a variável resposta sejam quantitativos, com suposição de normalidade dos resíduos. Pelo fato das variáveis possuírem mais de uma categoria e em ordem natural de avaliação, faz-se uma regressão logística politômica ordinal [5]. Por meio desta, é possível obter as probabilidade de determinada amostra pertencer às categorias. O terceiro passo é fazer uma análise de discriminantes, de modo que a regressão logística é utilizada como a função classificadora. Por fim, é conduzida a técnica que faz o uso da curva ROC para avaliar a qualidade da previsão do modelo. Os resultados das técnicas utilizadas por Rios são apresentadas ao final desse capítulo

2.1 Avaliação de textura de arroz

Conforme citado anteriormente, a avaliação da textura de arroz pode ser realizada por meio da análise sensorial de textura, da análise instrumental de textura e de medidas viscoamilográficas. O estudo realizado por Rios explora alternativas para a predição de textura via análise sensorial, com base em medidas instrumentais de textura e viscosidade, otimizando o processo de análise sensorial utilizada até então pela Embrapa.

Entre as alternativas para a realização do estudo de textura de arroz, a análise sensorial é considerada a ideal. Entretanto, todos os distintos processos para a análise do arroz possuem vantagens e desvantagens. Apesar da análise sensorial ser considerada como o método padrão, essa análise requer que profissionais sejam bem treinados, gerando custos adicionais e à necessidade de mais tempo para a avaliação. Além disso, esse tipo de

avaliação frequentemente gera dúvida no avaliador no que se refere a definição da categoria à qual pertence determinada amostra de arroz. A utilização de medidas instrumentais para prever a avaliação sensorial traz vantagens, pois a avaliação da categoria da amostra é definida por uma máquina, de modo que o processo não gera dúvida como a que é gerada em razão da subjetividade do avaliador. Em contrapartida, o custo de tal tecnologia é elevado e a obtenção de resultados é demorada. Por fim, a previsão da análise sensorial via medidas instrumentais de viscosidade torna a definição da categoria de maneira rápida e barata. Assim, a previsão da textura de arroz por meio de medidas viscoamilográficas é a mais vantajosa, otimizando o processo em vários sentidos. Neste estudo, estamos procurando melhorar a qualidade da previsão do modelo e informação fornecida sobre a incerteza envolvida na classificação por meio do modelo logístico de predição.

Este trabalho concentra atenção na avaliação de textura de arroz por meio de medidas de viscosidade e com foco na variável pegajosidade, para os dados de 2014 fornecidos pela Embrapa. Por meio do estudo realizado por Rios, foi possível notar uma baixa qualidade dos dados para a variável sensorial da dureza, que estão sendo revistos por técnicos da Embrapa e serão incluídos em estudos futuros.

2.1.1 Medidas de perfil viscoamilográfico

Aproximadamente 95% da matéria-prima do arroz é composta por amido, o que torna o torna4 o componente mais explorado no estudo do arroz. A tecnologia para extração do amido deste grão é elevado, de modo que alternativas para essa análise são fundamentais. Grande parte das propriedades físicas do arroz são explicadas pela amilose [4], de modo que suas características são fundamentais para este estudo. Sendo assim, a presente tecnologia prevê formas de obter informações sobre características da amilose de maneira rápida e barata, o que a torna uma alternativa para a análise embasada no conteúdo de amido, que apresenta custo elevado.

O início do processo de obtenção das medidas de interesse se dá pela colheita, debulha e secagem natural dos grãos de arroz. Logo em seguida, se faz o processamento da amostragem dos grãos. E, por fim, pode-se dar início ao processo de obtenção das informações necessárias [1].

No banco de dados fornecido pela Embrapa, sete variáveis fornecem informações sobre viscosidade e teor de amilose aparente do grão de arroz. Entre as medidas de

viscosidade, está o teor de amilose aparente dos grãos, representado pela variável TAAFIA. Essa variável é medida de maneira simples por meio da máquina mostrada na Figura 2.1.

Figura 2.1: Intrumento utilizado para medição da variável TAAFIA (teor de amilose aparente do arroz).

Outra medida de viscosidade é a variável TAASEC, que indica o teor de amilose absoluto dos grãos. As medidas para essa variável são obtidas por meio de um intrumento de manuseio mais complexo e de tecnologia mais cara, mostrada na Figura 2.2.

Figura 2.2: Intrumento utilizado para medição da variável TAASEC (teor de amilose absoluto do arroz).

Por fim, a medida de viscosidade TG representa a temperatura de gelatinização do grão. Para a obtenção dessa medida, o teste de dispersão alcalina é conduzido em placas

de plástico, tampadas e incubadas em uma estufa (Figura 2.3).

Figura 2.3: Intrumentos utilizados para medição da variável TG (temperatura de gelatinização).

Para a obtenção de cada uma dessas medidas de viscosidade, a amostra de arroz é colocada sob condições controladas específicas, reagindo com outros produtos a determinadas temperaturas. Quanto às medidas de perfil viscoamilográfico, há cinco variáveis de interesse, que estão ilustrados na Figura 2.4, em que quatro estão no banco de dados. As variáveis são [1]:

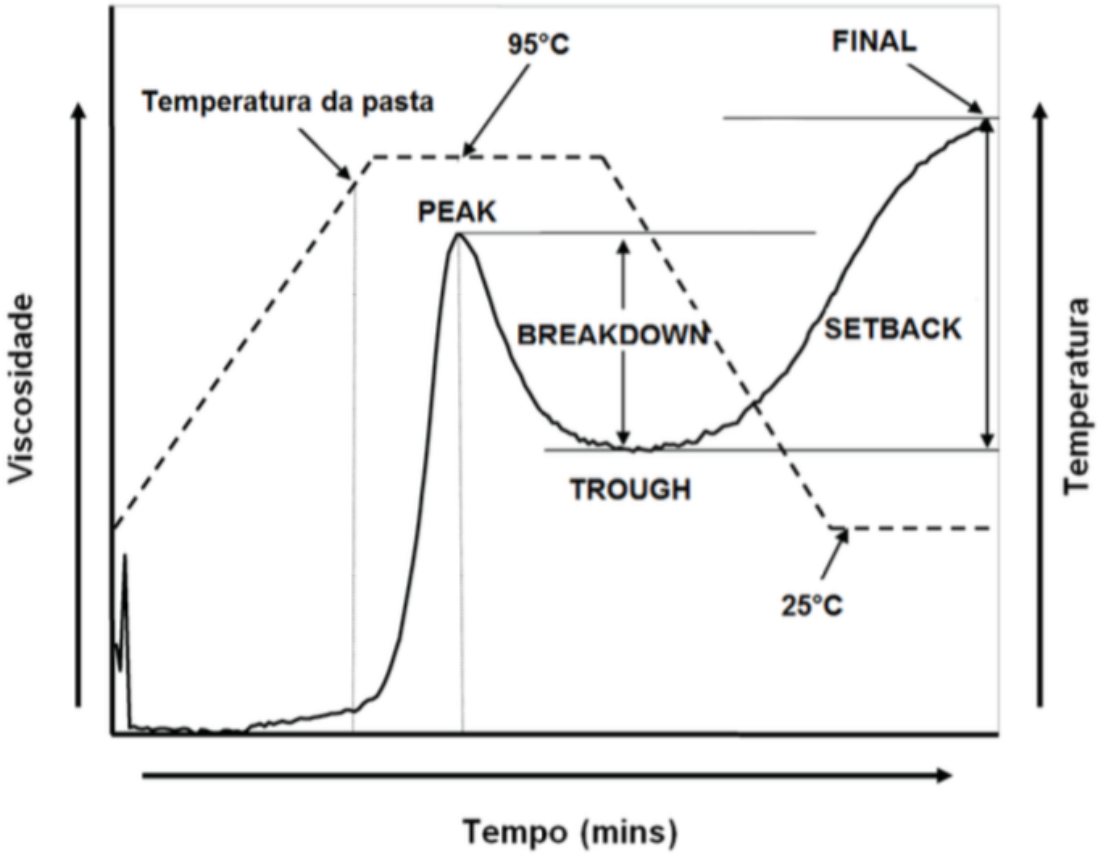

Figura 2.4: Medidas de perfil viscoamilográfico.

- PEAK (viscosidade de pasta mínima): maior valor da viscosidade durante o ciclo de aquecimento, que é obtido no ponto máximo da curva da figura;
- TROUGH (viscosidade de pasta mínima à quente): é o menor valor da viscosidade durante os 3 minutos em que a temperatura é mantida constante a 95°C;
- BREAKDOWN (quebra de viscosidade): diferença entre a viscosidade de pasta máxima e a viscosidade de pasta minima à quente;
- FINAL (viscosidade final): valor final da viscosidade durante o ciclo de resfriamento, que se dá a 25°C;
- SETBACK (tendência à retrogradação): diferença entre viscosidade final e a viscosidade de pasta minima à quente.

Essas medidas são obtidas por meio de um instrumento chamado de Texturômetro, representado na Figura 2.5. Tal tecnologia possui um custo mais vantajoso em relação ao

instrumento que faz a classificação sensorial do arroz de forma direta (o arroz é colocado na máquina e o instrumento já retorna a classe que a amostra deve ser alocada).

Figura 2.5: Equipamento utilizado para obtenção de medidas de perfil viscoamilográfico.

Conforme mostrado na Figura 2.5, a amostra do grão é colocada no equipamente e uma agulha entra em contato com a amotra e assim, são determinadas as medidas de interesse. A imagem demonstra um exemplo realizado com o feijão, entretanto o mesmo procedimento é conduzido com o arroz.

O objetivo deste estudo é fazer a substituição de técnicas e tecnologias de alto custo para a avaliação sensorial do arroz, buscando a otimização desse processo. Assim, as medidas utilizadas para a criação do modelo são as variáveis TAAFIA, TG, PEAK, BREAKDOWN e FINAL. Resultados semelhantes para as variáveis TAAFIA e TAASEC foram obtidos [1], assim a variável TAAFIA foi mantida, uma vez que a tecnologia para a obtenção dessa medida possui um custo mais vatanjoso em relação à tecnologia utilizada para a obtenção da variável TAASEC. As medidas TG, PEAK, BREAKDOWN e FINAL também são obtidas por meio de tecnologias de custo mais baixo.

As seções seguintes são uma breve revisão das técnicas utilizadas por Rios para a criação do modelo. Maiores detalhes podem ser obtidos em Rios (2015) [1], Agresti (2002) [5], Kleinbaum (2010) [6] e Weisberg (2005) [10].

2.2 Componentes Principais

A análise de componentes principais é utilizada a fim de permitir a redução de dimensão deste estudo com relação às variáveis de viscosidade. Essa análise é realizada por meio da explicação da matriz de variância-covariância dessas variáveis, através de suas combinações lineares.

A partir da matrix $X_{n\times m}$, em que n representa as linhas da matriz de uma mesma observação (neste caso uma mesma amostra de arroz) e m representa as colunas com as diferentes variáveis (nesse caso, de medidas de perfil viscoamilográfico), pode-se obter a matriz de variância-covariância (Σ), necessária para prosseguir com o cálculo. Por meio da matriz Σ, é possível obter os valores que explicam a dispersão entre os valores de uma variável e entre os pares de valores entre duas variáveis [1]. O próximo passo é encontrar a matriz de correlação, a matriz simétrica R, a partir da matriz de variância-covariância $\sum$, em que a diagonal da matriz R representa as variâncias, e o restante, as covariâncias. O intuito dessa nova matriz R é padronizar os dados das diferentes variáveis, obtidas com diferentes mensurações, de modo que a comparação direta entre os pares de variáveis se torne possível [1], realizando-se assim, a correlação entre as variáveis de perfil viscoamilográfico.

Para a obtenção das componentes principais, se faz necessário o cálculo dos autovalores de cada variável, representado pelo vetor lambda, em que a seguinte equação deve ser satisfeita,

$$|R - \lambda I| = 0, \tag{2.1}$$

em que

$|\cdot|$ = determinante

I = matriz identidade.

Em seguida, são encontradas as combinações lineares para cada variável por meio da equação 2.2 a seguir,

$$Y_p = a_{pi}^{'} X = a_{p1}X_{1p} + a_{p2}X_{2p} + \cdots + a_{pp}X_{pp}, \tag{2.2}$$

em que p = 1, 2, $\cdots$, m.

Por fim, as componentes principais se dão pelas combinações lineares que maximi-

zam a variância de Y_p [9, p. 431].

As componentes principais permitem a redução da dimensão do estudo por meio da manutenção das componentes principais responsáveis pela maior parte da variação total presente nos dados.

2.3 Regressão Logística

A análise de regressão busca definir a relação entre duas ou mais variáveis, de modo que a variável resposta possa ser prevista pelas variáveis consideradas explicativas. Dessa maneira, a criação de um modelo estatístico é desejável a fim de representar, da melhor forma possível, a relação dessa variável resposta com as demais. De maneira geral, os modelos estatísticos são compostos por três partes [5, p. 116]:

- Componente aleatória: consiste na variável resposta de uma distribuição de probabilidade da família natural dos exponenciais;

  Neste estudo, a variável resposta se dá por meio de medidas de textura sensorial. Estas são representadas por meio da distribuição multinomial, uma vez que as possíveis categorias de classificação são maiores que duas.

- Componente sistemática: define as possíveis variáveis associadas com a variável resposta de interesse, denominadas covariáveis ou variáveis explicativas.

  Neste caso, as variáveis a serem analisadas como possíveis variáveis associadas à variável de interesse (medidas de textura sensorial), são as variáveis de perfil viscoamilografico descritas anteriormente (TAAFIA, TAASEC, PEAK, etc).

- Função de ligação: especifica a função da média da variável resposta que equaciona a componente sistemática.

  Neste estudo, a função de ligação é a função logito, dada por

$$g[\pi(y)] = \log\left(\frac{\pi(y)}{1-\pi(y)}\right). \tag{2.3}$$

A fim de se verificar quais covariáveis no modelo possuem efeito significativo na estimação da variável resposta, utiliza-se o teste de Wald [1, 5, p.84]. As hipóteses para este teste são definidas a seguir:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0. \end{cases} \tag{2.4}$$

As hipóteses podem ser interpretadas como:

$$\begin{cases} H_0 : \text{covariável } x_i \text{ não exerce influência significativa no modelo} \\ H_1 : \text{ covariável } x_i \text{ exerce influencia significativa no modelo.} \end{cases} \tag{2.5}$$

O modelo de regressão logística pode ser expresso pela Equação 2.6.

$$\log\left(\frac{\pi(y)}{1-\pi(y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m \tag{2.6}$$

O software R utiliza um preditor linear alternativo, que será utilizado neste trabalho, que é dado por:

$$\log\left(\frac{\pi(y)}{1-\pi(y)}\right) = \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_m X_m. \tag{2.7}$$

Após conduzir a análise de componentes principais e feita a devida redução de variáveis, nem todas as m covariáveis serão utilizadas no modelo.

Por fim, a probabilidade estimada de se obter determinada característica (possíveis categorias em que a amostra pode ser classificada) se dá por meio da exponencializacão da função logito, dada pela equação 2.8 a seguir,

$$\hat{\pi}(y) = \frac{e^{\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_m x_m}}{1 + e^{\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_m x_m}}. \tag{2.8}$$

Para o caso da regressão linear, os betas seriam estimados através do método da máxima verossimilhança. Na regressão logística, por se tratar de um modelo não-linear, se faz necessário o uso do método numérico de estimação iterativa para a obtenção dessas estimativas, chamado método de Newton Raphson. O valor estimado é obtido por meio das soluções das equações que maximizam a função em questão [5, p. 143]. Além disso, o fato de ser um modelo não linear pode fazer com que a probabilidade calculada se dê em um valor maior que um. A fim de solucionar este problema, a função logito é utilizada.

2.3.1 Regressão Logística Politômica

Neste estudo, a principal medida sensorial de textura de arroz é a variável pegajosidade, que pode ser classificada em uma entre sete diferentes categorias. Essa variável segue uma distribuição multinomial, por possuir mais de uma categoria. Sendo assim, se faz necessário o uso da regressão logística politômica. Outro fator importante para a definição da técnica utilizada neste estudo, é o fato da variável pegajosidade ser ordinal (classificada de 1 a 7, em que 1 é extremamente solto, 2 é muito solto, e assim por diante). Há três tipos de modelos freqüentemente utilizados para o caso da regressão logística politômica ordinal [1, 11, p. 288-291]:

- Modelo de categoria adjacente: compara cada categoria k com a categoria anterior 1-k;
- Modelo de razão continua: compara cada categoria k com todas as categorias anteriores;
- Modelo de chances proporcionais: compara todas as categorias anteriores e equivalente à categoria k com todas as categorias acima.

O último tipo de modelo apresentado, também denominado de modelo logito cumulativo [1], é o modelo utilizado por Rios [1] e que será mantido neste projeto. Dessa maneira, após a análise de componentes principais, aplica-se a regressão logística nas componentes. O modelo utilizado é descrito pelas seguinte equações:

$$\log\left(\frac{\pi(y)}{1-\pi(y)}\right) = \beta_0 - \beta_1 E_1 - \beta_2 E_2 - \cdots - \beta_m E_m \tag{2.9}$$

e

$$\hat{\pi}(y) = \frac{e^{\beta_0 - \beta_1 E_1 - \beta_2 E_2 - \cdots - \beta_m E_m}}{1 + e^{\beta_0 - \beta_1 E_1 - \beta_2 E_2 - \cdots - \beta_m E_m}}, \tag{2.10}$$

em que E_i é o Escore i, que representa i-ésima componente principal, calculada previamente.

A probabilidade estimada da variável resposta de textura sensorial pertencer a categoria k, ou a uma categoria inferior a esta, é obtida via exponencialização da Equação 2.8, tornando possível o cálculo deste para diferentes valores da variável explicativa. Segue a fórmula utilizada para seu cálculo.

$$P(y \leq K|E_1, E_2, \cdots, E_m) = \frac{e^{\beta_{0k} - \beta_1 E_1 - \beta_2 E_2 - \cdots - \beta_m E_m}}{1 + e^{\beta_{0k} - \beta_1 E_1 - \beta_2 E_2 - \cdots - \beta_m E_m}}. \quad (2.11)$$

2.4 Classificação e discriminação

Segundo Wichern (2007) [9], a discriminação e a classificação são técnicas multivariadas focadas na separação de distintos grupos de observações, e na alocação de novas observações em grupos pré-determinados. A função que separa as observações geralmente é referida como o fator "alocador", e a regra que aloca essas observações pode ser referida como o procedimento discriminatório [9]. Geralmente os objetivos desse processo (separação e alocação) se sobrepõem e a distinção entre estes se torna difícil de ser identificada [9].

A regressão logistica é a função responsável pela separação das observações em grupos já pré-determinados. Neste estudo, queremos separar as diferentes amostras de arroz observadas, entre categorias de 1 a 7 (descritas anteriormente) já pré-estabelecidas pela Embrapa.

Um conceito importante para o estudo da classificação e discriminação é o conceito de valores preditos. Estes se dão pela probabilidade da observação i (amostra i) pertencer à categoria k [1], dado por $P(Y_i = k|E_{i1}, E_{i2}, \cdots, E_{im})$.

Por meio da Equação 2.11, a probabilidade da amostra i pertencer à categoria k é calculada, de modo que este é o critério utilizado para que as amostras sejam alocadas e separadas em diferentes grupos.

Para o caso da regressão logística politômica, possuímos K categorias, em que as observações podem ser classificadas, sendo que cada observação possuí K valores preditos, e há um valor para cada categoria K [1]. Sendo assim, entre esses K valores preditos, o maior valor indica a qual categoria K determinada observação pertence. Sendo assim, para o exemplo de 4 categorias, calcula-se

$$P(Y = 1|X) = P(Y \leq 1|X),$$
$$P(Y = 2|X) = P(Y \leq 2|X) - P(Y \leq 1|X),$$
$$P(Y = 3|X) = P(Y \leq 3|X) - P(Y \leq 2|X),$$
$$P(Y = 4|X) = 1 - P(Y \leq 2|X),$$

com base na Equação 2.11 apresentada anteriomente.

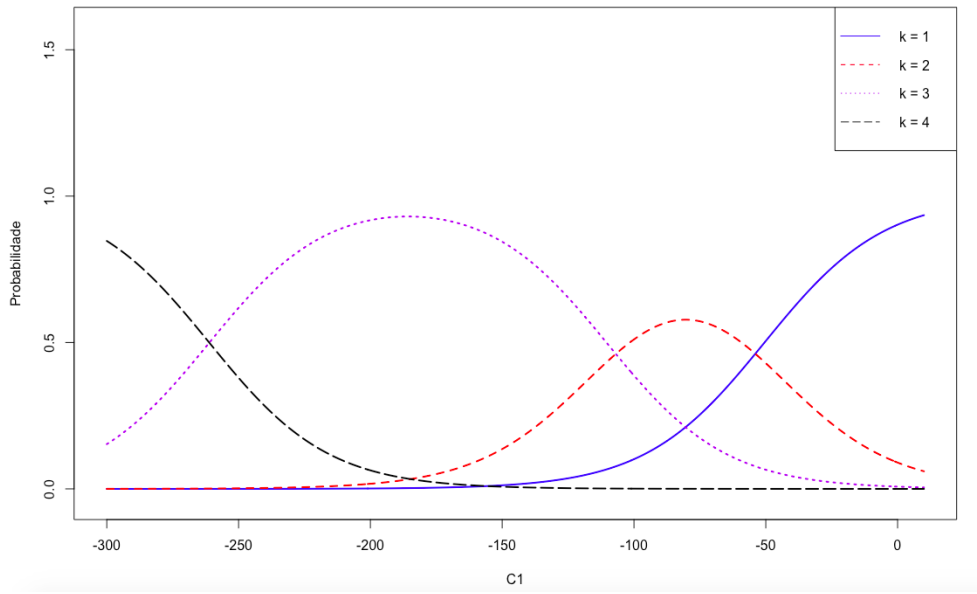

Figura 2.6: Probabilidade estimada de se obter a variável resposta pertencente a categoria k, considerando diferentes valores da variável explicativa X.

Cada curva na figura acima representa o valor predito para uma observação (amostra de arroz) de pertencer à K-ésima categoria. Observando a imagem, pode-se notar que a classificação não se dá de maneira tão simples. Em regiões de transição de categoria (Ex: de 2 para 3), onde ocorre o encontro das curvas, os valores preditos possuem resultados muito próximos, dificultando a alocação [1]. Ao contrário das regiões de transição, pode-se notar que para valores de pico, a escolha da categoria pode ser facilmente realizada.

Por fim, por meio da taxa de erro de classificação, pode ser medida a eficiência que a função de classificação tem em classificar as observações em suas verdadeiras categorias, uma vez que a verdadeira origem da população for conhecida [1]. Para o cálculo dessa medida é primeiramente conduzida a técnica de validação cruzada, e assim, a taxa é obtida por meio do erro aparente dos dados, que é calculado pela divisão de categorias classificadas erroneamente (que não correspondem à sua verdadeira população), sobre o número total de obsevações classificadas. Este cálculo resulta na porcentagem de observações classificadas erroneamente, ou seja, a taxa de erro de classificação.

2.4.1 Curva ROC

A fim de se acessar o quão acurada é a previsão do modelo utilizado, a ferramenta da curva ROC (Receiver Operating Characteristic) é utilizada. Dessa maneira, a eficiência preditiva do modelo utilizado para classificar as observações é avaliada por meio de uma

representação gráfica [6].

A seguir, serão descritos alguns conceitos teóricos fundamentais para a criação da curva ROC. Esse método busca encontrar a curva ROC e a área embaixo desta curva, denominada AUC (Area Under the Curve). A AUC mede a discriminação, que é a habilidade do modelo de classificar corretamente as observações pertencentes a "categoria principal" e a "categoria secundária" [1]. Entende-se como categoria principal, a verdadeira categoria em que a observação deve ser classificada, e como categoria secundária, a categoria em que a observação pode ter sido possivelmente classificada de maneira errônea.

A análise da qualidade da discriminação do modelo, ou seja, a curva ROC, é uma representação da Sensibilidade (Se) contra Especificidade (Es), duas medidas fundamentais para esta análise. Outra medida utilizada para a construção da curva ROC é a medida [1-(Es)] [6].

O modelo é considerado um bom preditor para suas discriminantes, quando a proporção de observações classificadas corretamente ("categoria principal") é maior que a proporção de casos classificados erroneamente ("categoria secundária"). Isso significa que o modelo é bom quando (Se) é maior que [1-(Es)] [6]. Idealmente, (Se) e (Es) deveriam ser iguais a 1 e [1-(Es)] deveria ser igual a 0.

De maneira resumida, o primeiro passo para esse processo está em encontrar a curva ROC, que se dá por meio de medidas de sensibilidade e especificidade [6]. Ao obtermos a curva ROC, podemos obter a AUC. Por fim, são utilizadas notas de cortes baseadas na AUC, buscando fazer uma classificação para a avaliação das discriminantes, representadas a seguir [1]:

- discriminação excelente: AUC de 0.9 até 1;
- discriminação boa: AUC de 0.8 até 0.9;
- discriminação razoável: AUC de 0.7 até 0.8;
- discriminação ruim: AUC de 0.6 até 0.7;
- discriminação péssima: AUC de 0.5 até 0.6;
- discriminação negativa: AUC de 0.0 até 0.5.

Para o caso da regressão logística politômica, que é o caso deste estudo, uma curva ROC para cada categoria K, em que o arroz pode ser classificado, é obtida. A curva

que possuir maior AUC representa a categoria que o modelo está prevendo com maior eficiência quando comparado com as outras categorias [1]. Um exemplo de curva ROC, quando se tem, por exemplo, 4 categorias (K = 4), é ilustrada na Figura 2.7.

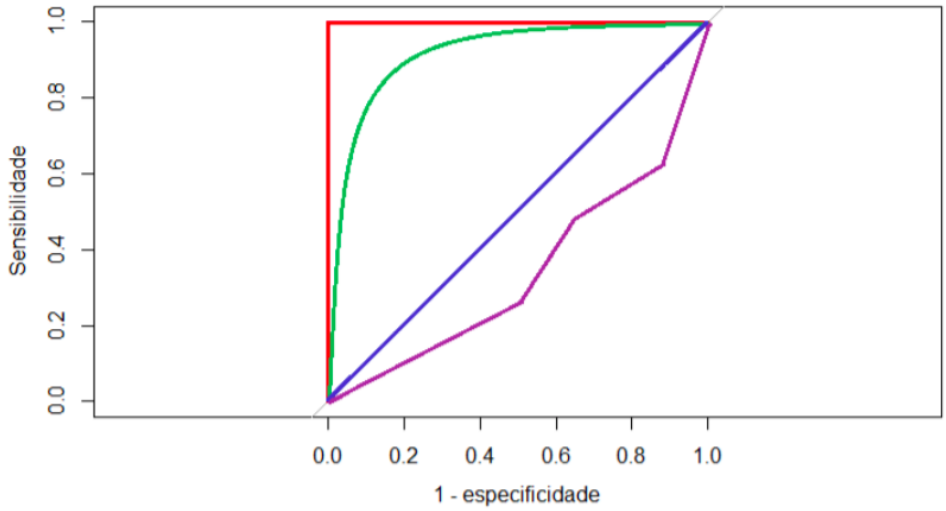

Figura 2.7: Exemplos de curva ROC.
Nota: Baseado em Rios (2015) [1]

Dessa maneira, a análise das curvas ROC para cada categoria são avaliadas a partir do gráfico e das classificações de avaliações apresentados acima. A curva vermelha na Figura 2.7 representa o caso em que a AUC resulta no valor 1, sendo esse o caso ideal em que a discriminação é excelente. A reta azul mostra um caso em que a AUC é de 0.5, ou seja, tem discriminação péssima. Para o caso da linha roxa, o seu valor se dá entre 0 e 0.5, resultando em uma discriminação negativa, em que o modelo faz a previsão de mais observações errôneas do que o número de observações corretamente previstas [1].

2.4.2 Resultados

Os resultados obtidos após o estudo conduzido por Rios [1] são fundamentais para o completo entendimento e realização desse trabalho. Dessa maneira, os resultados obtidos serão sintetizados nessa seção [1].

Após a análise de componentes principais e feita a redução de variáveis, foi possível observar que para os dados de terras altas, as duas primeiras componentes principais explicam 90,22% da variância total dos dados. Para o caso dos dados de terrenos irrigados, 86,21% da variância total é também explicada pelas duas primeira componentes principais.

Por meio da regressão logística, foi possível concluir que, para as terras altas, apenas a primeira componente exerce influencia significativa no modelo. Para os terrenos irrigados, manteve-se que as duas primeiras componentes exercem influencia significativa no modelo.

As probabilidades para cada categoria são obtidas por meio da regressão logística, sendo esta utilizada como função classificadora. Assim, a curva ROC é utilizada como um mecanismo gráfico auxiliar para a avaliação da acurácia da previsão do modelo. Para os dados de terras altas, as curvas de probabilidades obtidas são representadas na Figura 2.8.

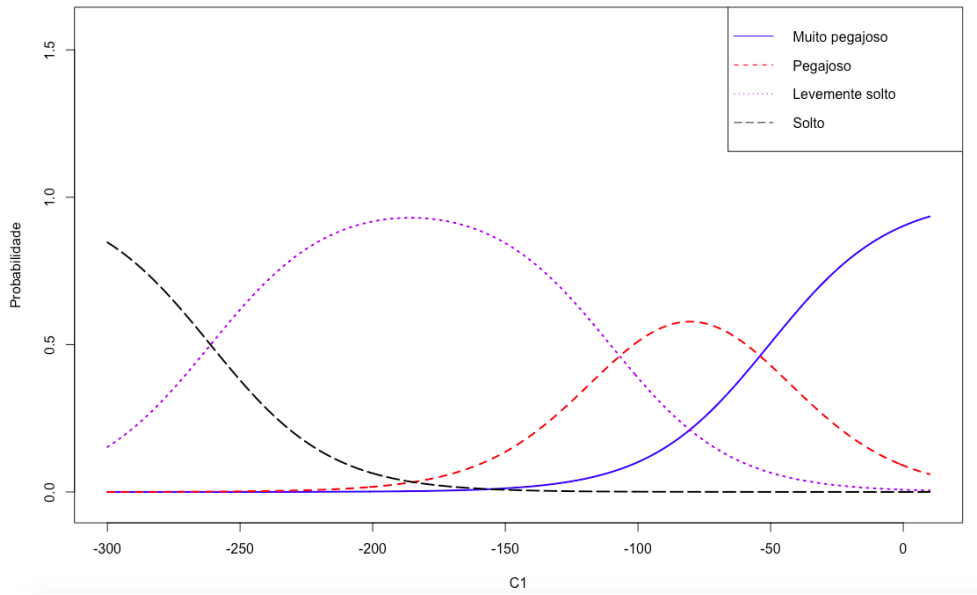

Figura 2.8: Probabilidades das categorias de avaliacao sensorial de pegajosidade, considerando diferentes valores da variavel C1, de arroz de Terras Altas para o ano de 2014 [1].

Por meio da validação cruzada, foi possível observar uma taxa de, aproximadamente, 23% de erro na predição da classificação das amostras de arroz [1]. Por fim, a Figura 2.9 apresenta o resultado da curva ROC para os dados de terras altas.

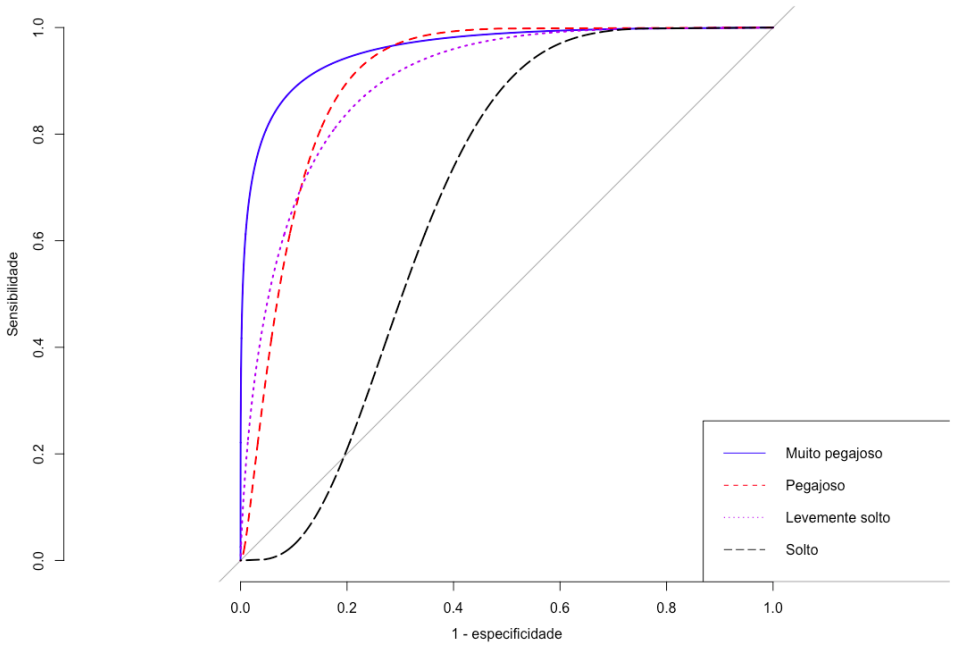

Figura 2.9: Curva de classificação ROC da avaliação sensorial de pegajosidade para o ano de 2014, prevista por meio do modelo de regressão logística, utilizando componentes principais de arroz de Terras Altas [1].

Com base na curva ROC apresentada, pode-se observar que a categoria Muito Pegajoso (MP) está sendo discriminada com maior precisão, com base no modelo utilizado e criado por Rios, do que as categorias Pegajoso (P) e Levemente Solto (LS) [1]. A categoria Solto (S) apresenta uma discriminação pobre, uma vez que apenas três amostras de arroz foram classificadas sensorialmente como tal, prejudicando a precisão da classificação nessa categoria .

Para os dados de terrenos irrigados, as curvas de probabilidade são representadas por superfícies, uma vez que são utilizadas duas componentes. As superfícies obtidas por meio do modelo são apresentadas na Figura 2.10.

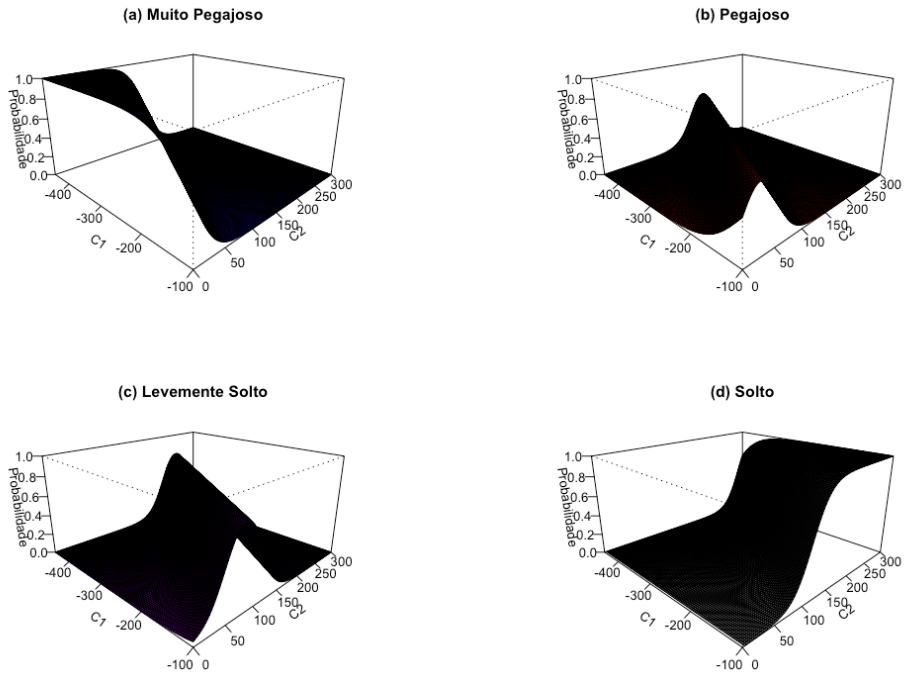

Figura 2.10: Probabilidades das categorias de avaliação sensorial de pegajosidade considerando diferentes valores das variáveis C1 e C2 de arroz para terrenos irrigados para o ano de 2014 [1].

Além disso, foi possível calcular a taxa de erro de classificação para as amostras de arroz para os dados em questão, resultando no valor de, aproximdamente, 37% de erro. Por fim, a seguir, é apresentada a curva ROC para os dados de terrenos irrigados, com base nos dados de 2014, utilizado por Rios.

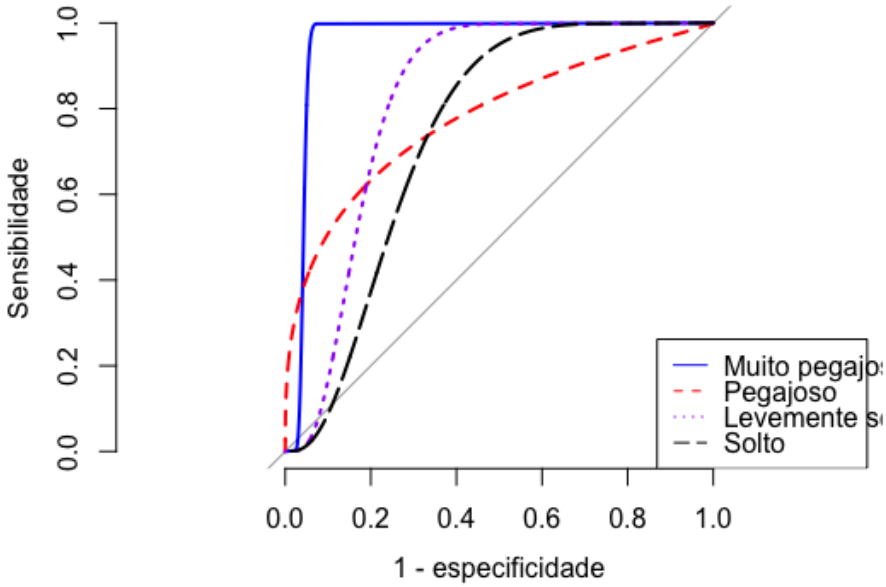

Figura 2.11: Curva de classificação ROC da avaliação sensorial de pegajosidade para o ano de 2014, prevista por meio do modelo de regressão logística utilizando a pegajosidade instrumental de arroz de Terrenos Irrigados [1].

2.5 Problemas das regiões de incerteza

Conforme citado anteriormente, as amostras de arroz podem ser classificadas em categorias de 1 a 7. Durante o estudo de Rios, notou-se que categorias extremas possuíam frequência muito baixa de observações, exigindo a redução do número de pontos na escala, para criação do modelo em quatro categorias. Dessa forma, nota-se uma falha na definição das escalas de mensuração inicialmente utilizadas para o estudo, afetando os resultados.

Outro problema identificado durante o estudo realizado por Rios, é a incerteza gerada ao se classificar as amostras de arroz. Há dois cenários em que a incerteza na escolha da escala da textura do arroz pode ocorrer. A primeira se dá pela incerteza do analisador, em que grãos com textura muito semelhante a outros possam ter sido colocados em categorias vizinhas, em como, por exemplo, 4 e 5. O segundo tipo de incerteza surge do próprio modelo criado por Rios, em decorrência das escalas originalmente utilizadas. Nesse caso, a definição da categoria associada a uma amostra de arroz, se dá por meio do cálculo da probabilidade do grão pertencer as distintas categorias, sendo a que possuir maior probabilidade a classificação escolhida pelo modelo. Em alguns casos, a probabilidade de duas categorias têm um valor muito próximo e uma dúvida na classificação também pode ocorrer. Esses dois casos representam o que é chamado, no estudo, de regiões de incerteza.

Neste trabalho, as regiões de incerteza serão estabelecidas, a fim de aumentar o acerto na escolha da categoria das amostras de arroz e, assim, reduzir o erro de classificação. A Figura 2.12 servirá de auxilio para a compreensão dessas regiões.

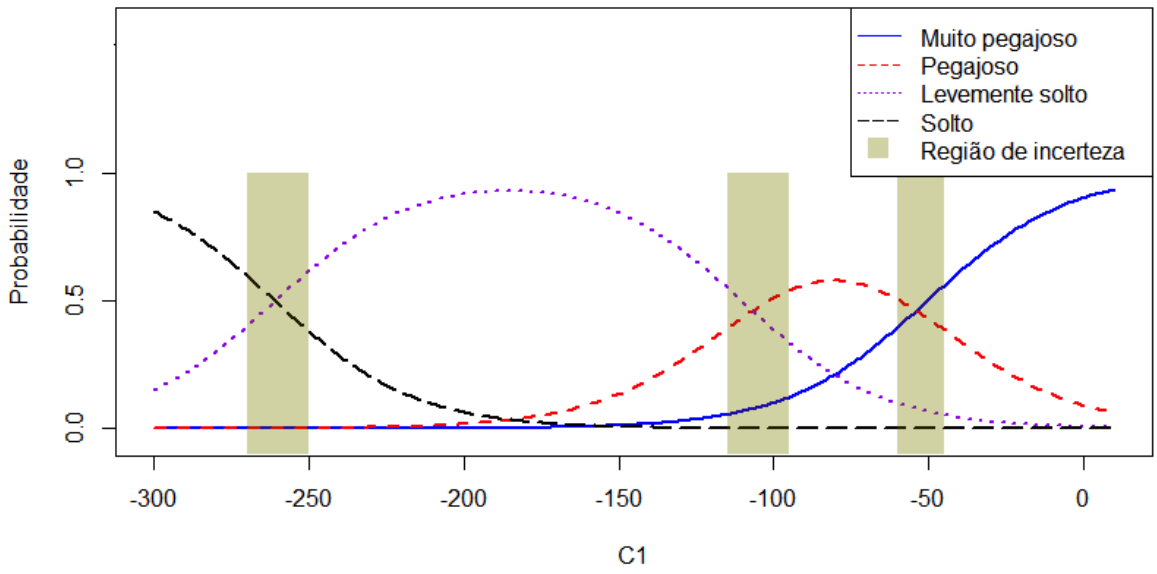

Figura 2.12: Probabilidades das categorias de avaliação sensorial de pegajosidade considerando diferentes valores da pegajosidade instrumental de arroz de Terras Altas para o ano de 2014 (Rios [1]).

A Figura 2.12 representa valores da primeira componente principal, formada pelas variáveis de perfil viscoamilográfico, que está indicada por "C1". As regiões de incerteza são apenas ilustrativas, e nenhum cálculo foi conduzido a fim de se determinar os limites destas. Em alguns casos, essa definição pode ser feita com extrema facilidade, como ilustrada na Figura 2.13.

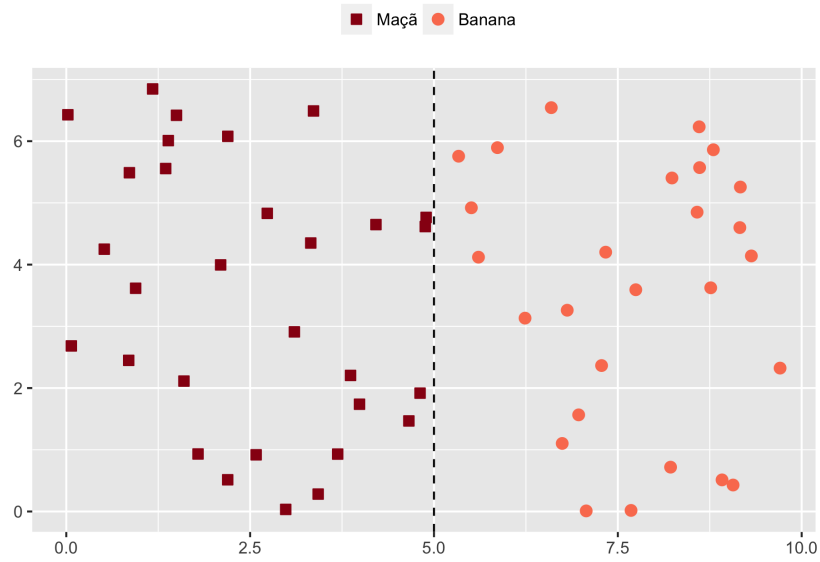

Figura 2.13: Exemplo de escolha de divisão de dados.
Nota: Baseado em Hartshorn, 2016 [12]

Neste exemplo, cada forma representa uma amostra pertencente a determinada categoria. Dessa maneira, é visível onde poderíamos traçar a reta que distingue as duas categorias. Em alguns casos, a distinção destas não é de fácil percepção, como ilustrado na Figura 2.14.

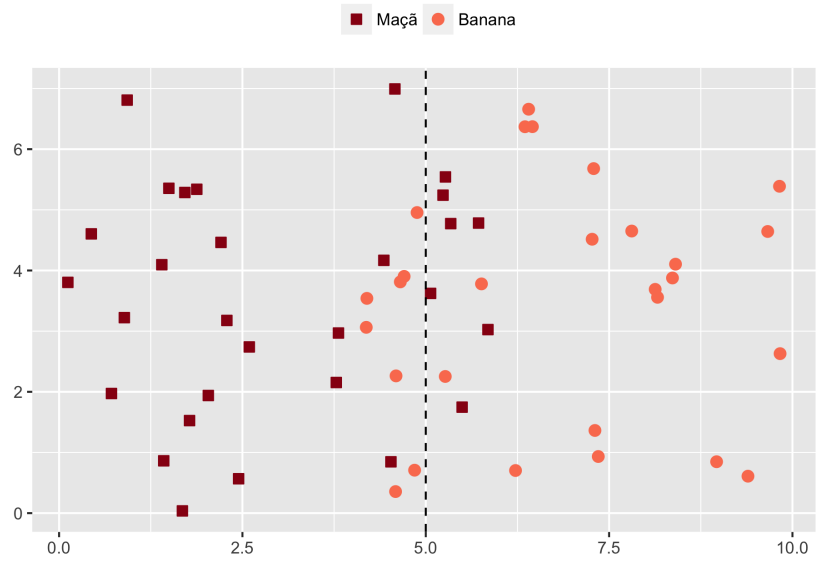

Figura 2.14: Exemplo de escolha de divisão de dados [12].

Neste caso, pode-se observar que ocorre a presença de erros cometidos quanto a escolha da escala da amostra. Na maioria das vezes, esses erros ocorrem na barreira de

transição entre as categorias, em que determinadas amostras deveriam ter sido classificadas como uma, mas foram classificadas como outra. Quando isso ocorre, a definição da reta de distinção se torna mais complexa, como será explicado com base nas Figuras 2.15, 2.16 e 2.17.

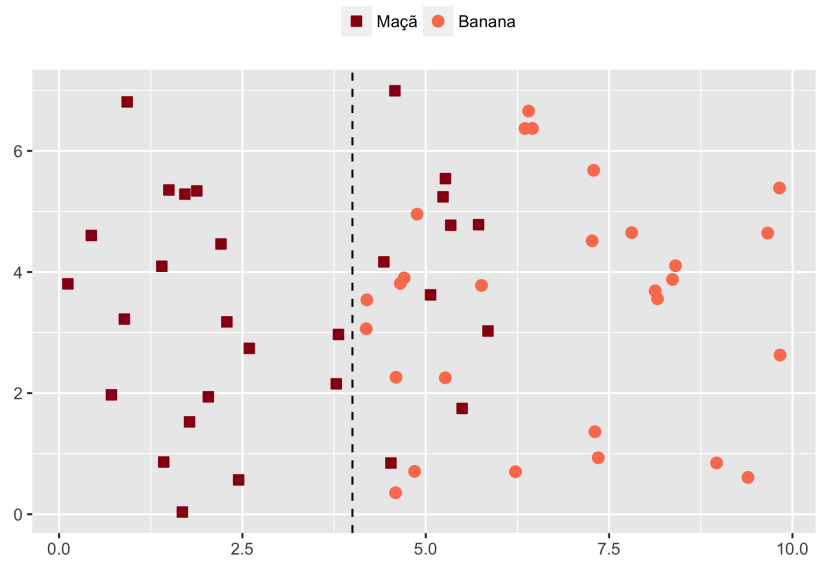

Figura 2.15: Exemplo de escolha de divisão de dados [12].

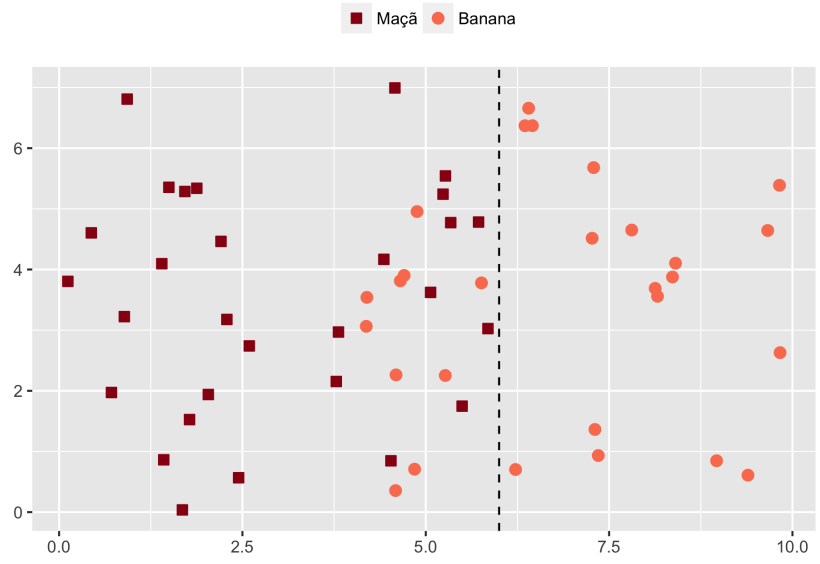

Figura 2.16: Exemplo de escolha de divisão de dados [12].

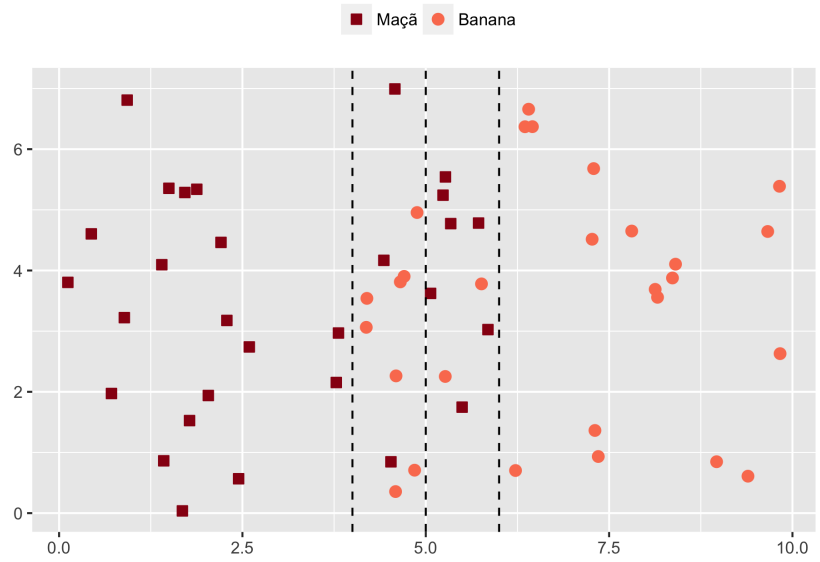

Figura 2.17: Exemplo de escolha de divisão de dados [12].

Se colocarmos uma divisão ao centro, ainda haverão amostras em categorias às quais elas não pertencem, representado pela Figura 2.14. Se movermos a barra para a esquerda, a fim de englobar a observação que foi classificada erroneamente, estaremos incluindo mais observações que não pertencem a essa categoria, aumentando o erro desta categoria, como pode-se observar na Figura 2.15. O mesmo ocorre se tentarmos mover a barra para a direita (Figura 2.16). Assim, são criadas as regiões de incerteza, compostas por dados que foram classificados erroneamente, conforme representado na Figura 2.12 e na Figura 2.17.

Este trabalho visa definir exatamente quais são os limites dessas regiões, minimizando o erro na escolha da categoria. Após a definição, serão implementadas técnicas auxiliares para a avaliação dessas barreiras, e como essas influenciam na classificação das amostras de arroz.

3 Metodologia

A base da dados utilizada neste trabalho foi fornecida pela Embrapa. Os dados foram disponibilizados em planilha Excel, lidos e analisados na versao 3.1.0 do software R, utilizando o ambiente de desenvolvimento integrado R-Studio. Para o desenvolvimento do presente trabalho, o completo entendimento e estudo do modelo criado por Rios [1] foi realizado, utilizando as técnicas de componentes principais, regressão logística, análise de discriminantes e curva ROC.

Para que o objetivo deste trabalho seja atingido, foram utilizadas técnicas de classificação de dados, tais como o Índice de Gini e o Critério de Entropia, para a construção das regiões de incerteza. Além disso, durante o desenvolvimente deste, foram empregados critérios para a definição de regiões de incerteza, até então não encontrados na literatura para este tipo de aplicação. Diversas técnicas são utilizadas com a finalidade de comparação entre elas, e como de se descobrir como afetam os resultados de interesse.

Após o estabelecimento das regiões de incerteza, são encontrados os erros de classificação para cada um dos casos. Assim, é possível fazer uma discussão acerca dos resultados obtidos e estabelecer uma medida de custo, com o objetivo de auxiliar o avaliador a medir o impacto de cada tipo de região de incerteza na classificação do arroz e, assim, tomar sua decisão quanto ao tipo de barreira que deseja empregar em sua análise. Por fim, outras medidas são utilizadas para a avaliação da qualidade das definição das regiões de incertezas, aqui chamadas de Erro Proporcional e Classificação Correta na Região de Incerteza.

3.1 Definição das regiões de incerteza

Neste estudo definimos regiões de incerteza com base em duas situações: (i) incerteza gerada por meio da classificação do avaliador e (ii) incerteza gerada pelo modelo criado por Rios. Originalmente os dados de arroz foram classificados em sete possíveis categorias. Conforme citado anteriormente, devido à baixa frequência de dados em determinadas categorias, estas foram reduzidas para apenas quatro: Solto, Levemente Solto, Pegajoso e Muito Pegajoso.

Durante a avaliação sensorial do arroz, é possível que a textura de uma amostra que deve ser classificada na categoria Solto, por exemplo, seja muito parecida com a textura

de uma amostra classificada na categoria vizinha, Levemente Solto. Dessa maneira, há uma incerteza na definição das categorias adjacentes por parte do avaliador, e assim a criação de regiões em que existe mais dificuldade de definição da categoria de classificação com base no modelo de predição utilizado.

De acordo com o modelo de previsão da escala sensorial do arroz [1], o critério classificador é definido com base na probabilidade de uma amostra pertencer a cada uma das categorias, sendo a escolha feita pela categoria que tiver maior probabilidade de ocorrência. Dessa maneira, a regra de decisão para a classificação entre categorias vizinhas se dá a partir do ponto de encontro das curvas de probabilidade destas, em que a transição das classes se encontra no ponto de probabilidade p_i, com $i = 1, 2, 3$, para cada transição entre categorias adjacentes. A Figura 3.1 ilustra um exemplo da barreira de transição entre as categorias Solto e Levemente Solto.

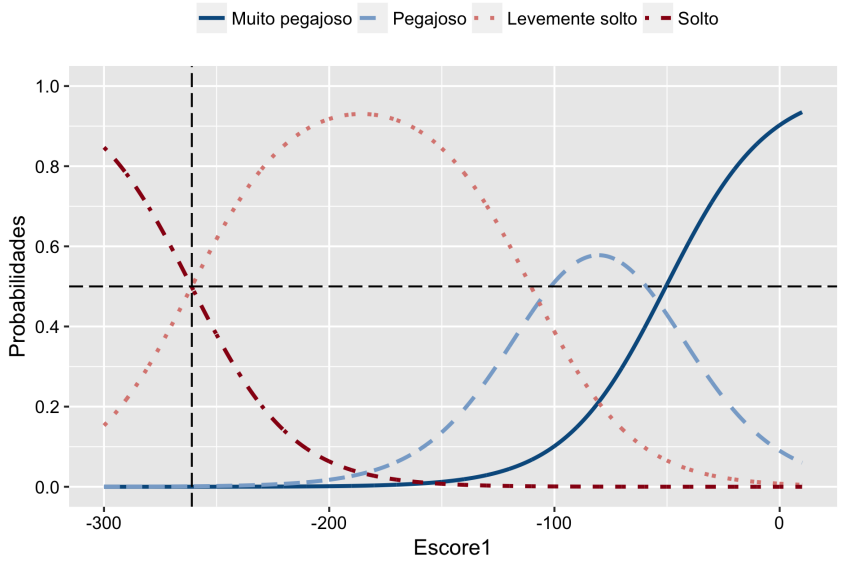

Figura 3.1: Barreira de transição de classificação entre categorias Solto e Levemente Solto, com base no modelo de Rios [1].

Conforme o exemplo, tudo o que está ao lado esquerdo do ponto p_1 de encontro 0,5, possui maior probabilidade na categoria Solto, sendo esta a categoria em que a amostra será alocada. O mesmo ocorre para a classificação da amostra para a categoria da direita, em que todas amostras possuem maior probabilidade na categoria Levemente Solto, sendo essa a categoria em que a amostra será alocada. A classificação do arroz para este caso é clara. Dessa maneira, é de interesse desse estudo estabelecer regiões de incerteza com base na categorização realizada pelo modelo, de modo que o critério de distinção da classificação

das amostras seja conduzido de maneira mais eficiente, eliminando as incertezas geradas a partir do modelo.

A definição desses intervalos ocorrem para dois casos, de acordo com os resultados de Rios: (i) para terras altas em que apenas a primeira componente principal é significativa, e (ii) para dados de terrenos irrigados em que as duas componentes principais são significativas. As quatro primeiras técnicas descritas a seguir (seções 3.1.1 e 3.1.2) servirão como base para a definição das barreiras de incerteza para os dados de terras altas. Um método adicional (seção 3.1.3) é utilizado para definição das regiões de incerteza para os dados de terrenos irrigados, associado com as quatro primeiras técnicas. Por fim, será realizada uma análise e comparação de resultados, de acordo com o tipo de definição de barreiras empregado.

3.1.1 Regiões de incerteza com base na classificação do avaliador

A. Barreiras Eliminatórias de regiões de confusão com base na dispersão real

O primeiro método para definição das regiões de incerteza com base na incerteza de classificação do avaliador, é realizado de maneira intuitiva a partir do gráfico de dispersão das amotras classificadas. Por meio da regressão logística, é possível obter os valores dos escores, para cada amostra de arroz. Assim, é possível obter o gráfico de dispersão dos dados com base nos escores obtidos, segundo a categoria definida por via sensorial pelo avaliador. A Figura 3.2 ilustra um exemplo do gráfico de dispersão descrito.

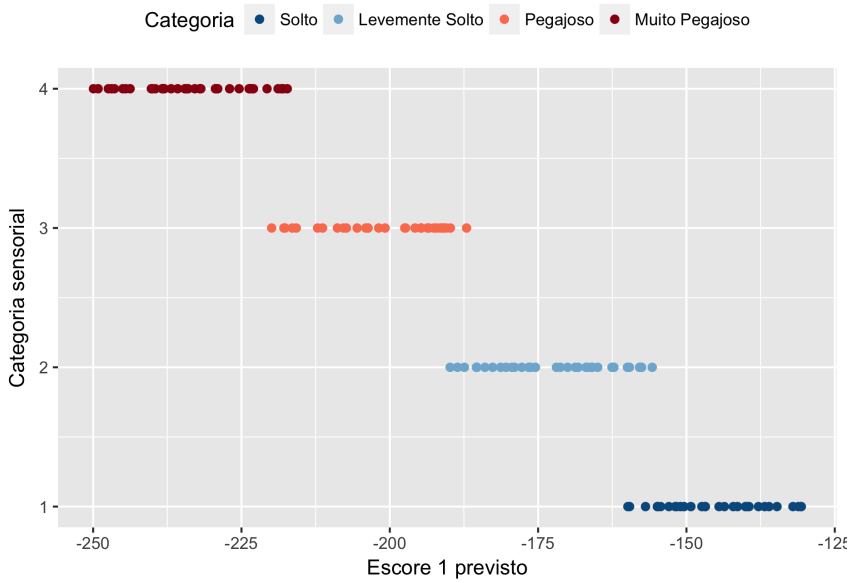

Figura 3.2: Gráfico de dispersão de exemplo de dados com base nos Escores1 preditos pelo modelo, segundo classificação original sensorial da amostra.

Por meio do gráfico acima, é possível observar que nas regiões de transições entre categorias, há uma incerteza gerada na classificação destas, uma vez que nessas transições alguns dados de categorias diferentes se sobrepõem. Sendo assim, há a criação de regiões de incerteza entre a transição das categorias, exemplificadas na Figura 3.3.

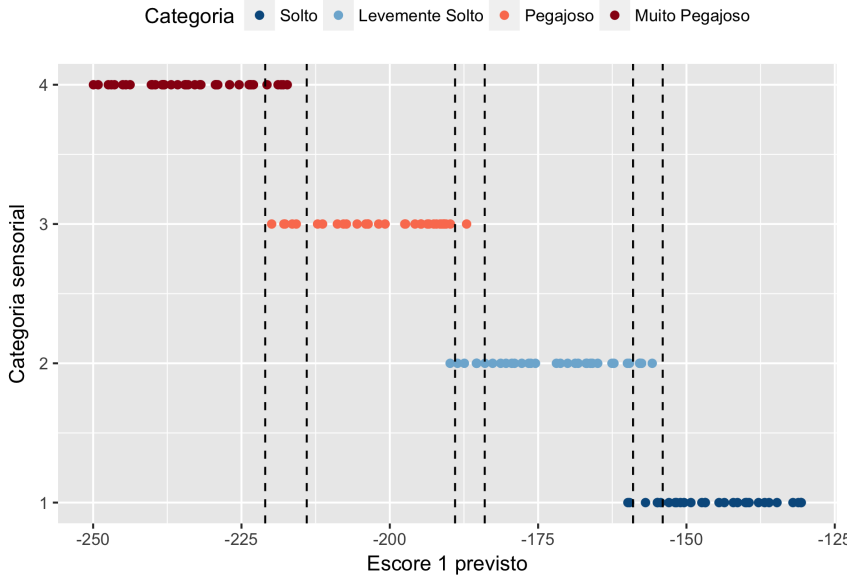

Figura 3.3: Regiões de incerteza com presença de sobreposição de dados de diferentes categorias.

É considerada como incerteza qualquer observação que cair entre os pontos que delimitam onde começa e onde termina a sobreposição de classificação entre categorias

distintas. Ou seja, a partir do ponto em que uma amostra se sobrepõe a outra que está classificada em uma categoria diferente, até o ponto em que essa sobreposição não é mais identificada, é considerada como uma incerteza, conforme representado na Figura 3.3. Por fim, após a definição das barreiras com base nos dados de dispersão, são plotadas as regiões de incerteza no gráfico com as curvas de probabilidades estimadas.

B. Barreiras com base em Medidas de Categorização de dados

O uso de medida originárias da Toria da Informação [12] pode ser útil para determinar a melhor divisão na separação dos dados.

B.1. Índice de Gini

O critério de Gini, também denominado de impureza de Gini, é uma das técnicas utilizadas para a classificação de dados, e é dado com o uso da equação

$$\text{Gini} = 1 - \sum_{j} p_j^2, \tag{3.1}$$

em que p_j é a probabilidade de uma amostra dos dados pertencer a uma determinada classe j. O valor ideal para esse índice é zero, e isso ocorre em uma situação em que sua amostra pertence 100% a uma determinada classe (Gini = 1 - 1 = 0) [12].

Considere uma situação em que existem quatro categorias possíveis. O cálculo do índice de Gini será,

$$\text{Gini} = 1 - \sum_{j=1}^{4} p_j = 1 - (p_1^2 + p_2^2 + p_3^2 + p_4^2). \tag{3.2}$$

e assim analisadas. Podemos chegar a uma conclusão sobre qual é a melhor divisão para os dados. Sendo assim, a divisão que possuir o menor valor do índice, será considerada a ideal. Dessa maneira, esse processo é realizado diversas vezes, buscando aproximar este resultado do valor zero, sendo esta a situação perfeita.

Os limites das regiões de incerteza são definidos por meio das duas melhores divisões encontradas, ou seja, pelas duas divisões que possuírem os menores valores para o Índice de Gini. Após conduzir todas as divisões possíveis, sabe-se que as demais divisões fora dessa região resultam em índices mais altos, não sendo as ideais. Por fim, de acordo com critério estabelecido, são plotadas as regiões encontradas no gráfico com as probabilidades estimadas.

B.2. Critério de Entropia

O processo do critério de entropia é muito semelhante ao de Gini. O cálculo da entropia é obtida por

$$\text{Entropia} = \sum_{j} -p_j \log_2(p_j), \tag{3.3}$$

de modo que a probabilidade, p_j, corresponde à chance de determinada amostra pertencer a uma categoria j. O alvo do critério de entropia é que seja atingido o valor zero, indicando que o conjunto de dados sendo analisados pertençam 100% a uma mesma categoria.

De maneira semelhante ao critério embasado no Gini, a decisão da melhor divisão para os dados é realizada com base no cálculo desse índice para todas as possíveis divisões, repetindo esse processo até que se chegue a um valor o mais próximo de zero possível [12]. A escolha das divisões que resultam nos limites das barreiras de incerteza, são definidas a partir das duas divisões que resultaram nos menores valores, quando for usado o Critério de Entropia. Feito isso, as divisões são plotadas no gráfico das probabilidades.

B.3. Exemplo

A fim de entender um pouco mais como funcionam os índices descritos, será realizado um exemplo simples para as duas técnicas, com dados gerados artificalmente, para três possíveis situações dos resultados dos índices: alto, baixo e moderado.

Em uma situação em que um processo de aprendizado para a classificação de frutas em que dispomos de 29 maçãs e 28 bananas (exemplo baseado em Hartshorn (2016) [12]), o vermelho representando as maçãs e o amarelo representando as bananas, na Figura 3.4

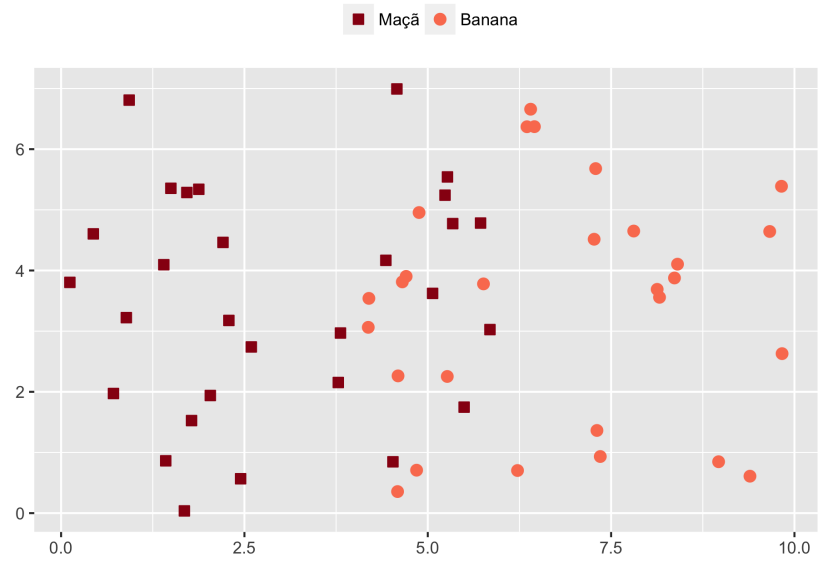

Figura 3.4: Dados para exemplo de maçãs e bananas [12].

Escolha de uma divisão - 1^a alternativa

A Tabela 3.1 a seguir apresenta informações sobre os dados que servirão de auxílio para o cálculo dos índices de Gini e a Entropia.

Tabela 3.1: Estimativas auxiliares para cálculo de índices.

Fruta	Frequência	Proporção (p_j)
Maçã	29	0,51
Banana	28	0,49
Total	57	1

Dessa maneira, o índice de gini e o critério de entropia são computados da seguinte maneira:

$$\text{Gini} = 1 - (0,51^2 + 0,49^2) = 0,4998,$$

$$\text{Entropia} = (-0,51)\log_2 0,51 - (0,49)\log_2 0,49 = 0.9997.$$

Escolha de duas divisões - 2^a alternativa

Para que possamos chegar a uma conclusão de qual é a melhor maneira possível de se de fazer essa divisão, diferentes alternativas para a criação deste grupos de dados são conduzidas. O resultado ótimo para a escolha da divisão ideal, é aquela que possuí o índice mais próximo de zero. Neste caso, a escolha da divisão separa os dados em dois grupos, conforme apresentado na Figura 3.5.

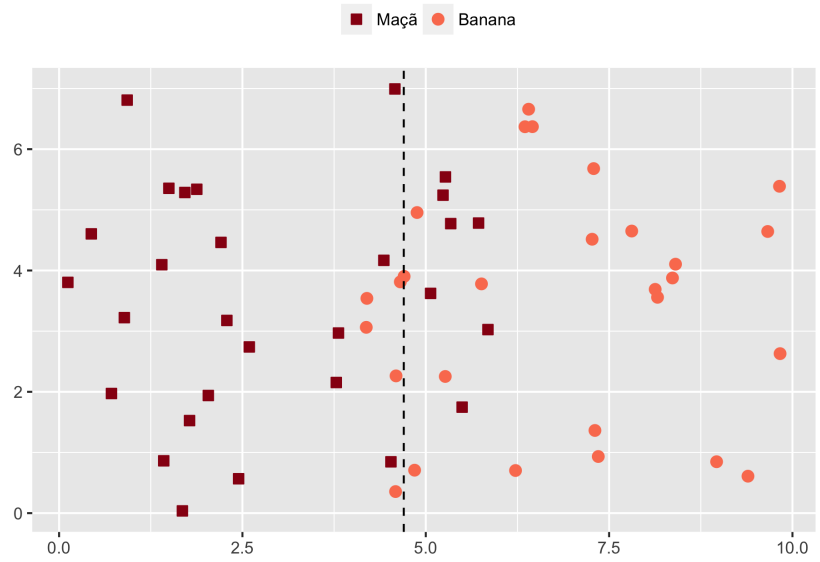

Figura 3.5: Exemplo de escolha de divisão de dados [12].

Cada grupo é composto pelas seguintes observações:

- Grupo 1: 22 maçãs e 5 bananas;
- Grupo 2: 7 maçãs e 23 bananas.

A Tabela auxiliar 3.2 apresenta as medidas necessárias para o cálculo dos índices, em que 1 representa o primeiro grupo, e 2 representa o segundo grupo.

Tabela 3.2: Estimativas auxiliares para cálculo de índices.

Divisão	Classe	Frequência	Proporção (p_j)
1	Maçã	22	0,81
1	Banana	5	0,19
2	Maçã	7	0,23
2	Banana	23	0,77

Dessa maneira, o Índice de Gini e o Critério de Entropia devem ser encontrados para os dois grupos, que é feito por meio das equações 17 a 20.

$$\text{Gini1} = 1 - (0,81^2 + 0,19^2) = 0,3078 \tag{3.4}$$

$$\text{Entropia1} = (-0,81)\log_2 0,81 - (0,19)\log_2 0,19 = 0,7015 \tag{3.5}$$

$$\text{Gini2} = 1 - (0,23^2 + 0,77^2) = 0,3542 \tag{3.6}$$

$$\text{Entropia2} = (-0,23)\log_2 0,23 - (0,77)\log_2 0,77 = 0,7780 \tag{3.7}$$

Por fim, para chegar ao resultado final de interesse, se faz necessária a realização de uma média ponderada pelo total de frutas em cada grupo, isto é,

$$\text{Gini Ponderado} = \frac{27\text{Gini1} + 30\text{Gini2}}{57} = 0,33 \tag{3.8}$$

$$\text{Entropia Ponderada} = \frac{27\text{Entropia1} + 30\text{Entropia2}}{57} = 0,7417 \tag{3.9}$$

Escolha de duas divisões - 3^a alternativa

Nesse caso, a escolha da divisão dos grupos se encontra próxima à margem esquerda do gráfico, de modo que é possível fazer uma comparação do comportamento dos índices em tal situação, conforme a Figura 3.6.

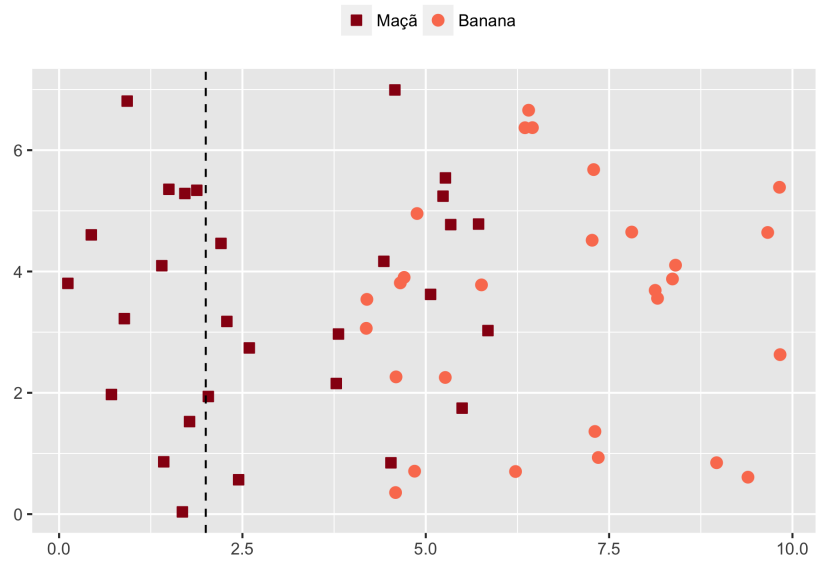

Figura 3.6: Exemplo de escolha de divisão de dados [12].

Neste caso cada grupo é composto por:

- Grupo 1: 12 maçãs e 0 bananas;
- Grupo 2: 17 maçãs e 28 bananas.

Além disso, é apresentada a tabela com medidas importantes para o cálculos dos índices.

Tabela 3.3: Estimativas auxiliares para cálculo de índices.

Divisão	Classe	Frequência	Proporção (p_j)
1	Maçã	12	1
1	Banana	0	0
2	Maçã	17	0,38
2	Banana	28	0,62

Dessa maneira, o Índice de Gini e o Critério de Entropia são encontrados por meio das equações 17 a 20.

$$\text{Gini1} = 1 - (1^2 + 0^2) = 0 \tag{3.10}$$

$$\text{Entropia1} = (-1)\log_2 1 = 0 \tag{3.11}$$

$$\text{Gini2} = 1 - (0,38^2 + 0,62^2) = 0,4712 \tag{3.12}$$

$$\text{Entropia2} = (-0,38)\log_2 0,38 - (0,62)\log_2 0,62 = 0,9580. \tag{3.13}$$

Dessa maneira, encontrados os valores dos Índices de Gini e de Entropia, calcula-se a média ponderada para tais estimativas, como demonstrado nas equações 21 e 22.

$$\text{Gini Ponderado} = \frac{12\text{Gini1} + 45\text{Gini2}}{57} = 0,3720 \tag{3.14}$$

$$\text{Entropia Ponderada} = \frac{12\text{Entropia1} + 45\text{Entropia2}}{57} = 0,7563. \tag{3.15}$$

Comparação

Por fim, após a realização das possíveis divisões do exemplo, é possível fazer uma comparação dos índices de acordo com as três separações realizadas. A Tabela 3.4 apresenta os valores encontrados para os dois critérios utilizados neste estudo, segundo o tipo de divisão utilizado.

Tabela 3.4: Comparação entre as três divisões escolhidas segundo índice utilizado.

Critério	Divisão 1	Divisão 2	Divisão 3
Gini	0,4998	0,33	0,372
Entropia	0,9997	0,7417	0,7563

A divisão que tiver o menor valor do índice do Gini, será considerada a melhor. Sendo assim, por meio da tabela é possível observar, que na primeira divisão, quando se faz apenas a divisão entre maçãs e bananas, foi encontrado o maior valor sendo a melhor entre as três escolhas. Pela observação dos resultados, nota-se que entre as três divisões a divisão 2 é a ideal.

A mesma conclusão é obtida com a Entropia, entretanto, nesta técnica é possível realizar o cálculo do ganho de informação de uma divisão para outra, que se dá por meio da subtração da entropia sem divisão (apenas diferenciando maçãs e bananas) pelo menor valor encontrado depois de feitas outras possíveis divisões (separação de dois grupos). Neste exemplo o cálculo se dá pela subtração dos índices de entropia das de duas divisões:

$$0{,}9997 - 0{,}7417 = 0{,}258.$$

Sendo assim, este é o maior ganho de informação para essas três divisões. De forma similar ao de Gini, a divisão 2 possuí o menor valor, sendo considerada a melhor dentre as três.

Ambos os índices, de Gini e de Entropia, possuem a mesma finalidade e cálculos muito semelhantes — assim, as análises para ambos os índices levam a resultados muito semelhantes.

Foram apresentados três exemplos para permitir o entendimento das técnicas de classificação; em que os valores dos índices desejados diminuiram ou aumentaram após executadas outras alternativas para separação dos dados em três cenários. O exemplo da categorização de maçãs e bananas foi reproduzido de um exemplo apresentado em Hartshorn (2016) [12], com diferentes dados.

Cálculo para mais de duas categorias

O problema das regiões de incerteza, apresentado na seção anterior, foi embasado em casos em que só há a classificação dos dados em duas categorias. Entretanto, este não é o caso do Projeto QualiArroz realizado pela Embrapa, em que após a redução de escalas, os dados foram alocados em quatro categorias. Para o caso em que há a presença

de mais de duas categorizações possíveis, a mesma incerteza será gerada para a todas as transições entre as categorias. Como este trabalho está focado nos dados relativos ao Projeto QualiArroz, é apresentada, a seguir, a forma do cálculo do índice de gini quando se têm quatro categorias, como mostrado no exemplo da Figura 3.7.

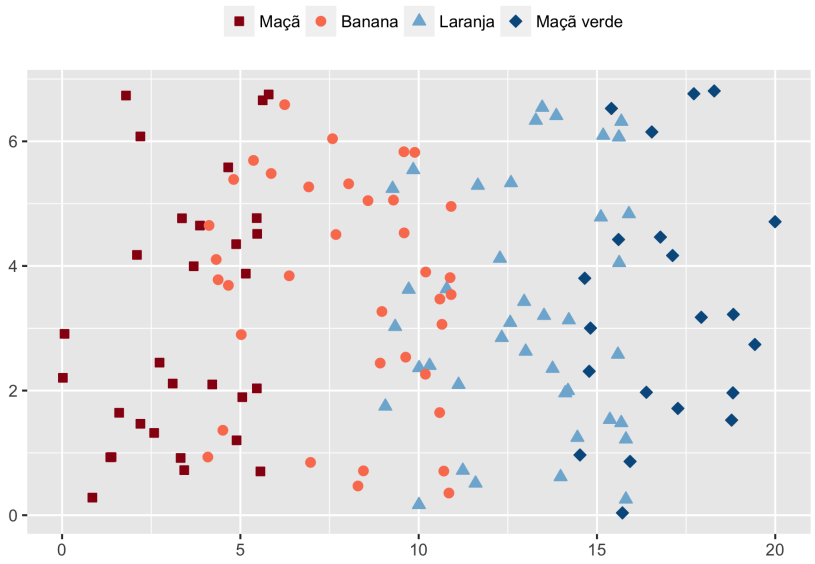

Figura 3.7: Exemplo com dados classificados em mais de duas categorias.

Analogamente ao caso de apenas duas categorias, ocorre o surgimento de mais de uma região de incerteza, ilustrado por meio do exemplo mostrado na Figura 3.8.

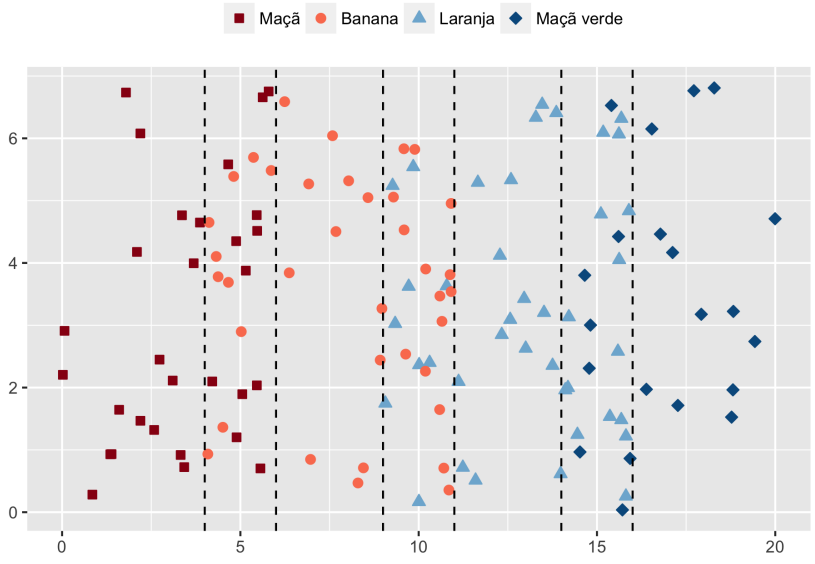

Figura 3.8: Regiões de incerteza para mais de duas categorias.

Nesse caso, os dados são tratados de maneira diferente do caso exemplificado ante-

riormente. Para o caso de quatro categorias, há três situações em que os índices de Gini e de Entropia devem ser calculados, ou seja, esse procedimento será realizado para as 3 barreiras de transição entre as categorias. Para a primeira situação, deseja-se encontrar a região de incerteza entre as categorias maçã e banana. O cálculo dos índices em questão é realizado de modo a se testar tudo o que está de um lado da divisão, contra tudo o que está do outro lado da divisão. Sendo assim, para o lado direito, tudo o que não for banana, passará a ser considerado como tal, conforme ilustrado na Figura 3.9.

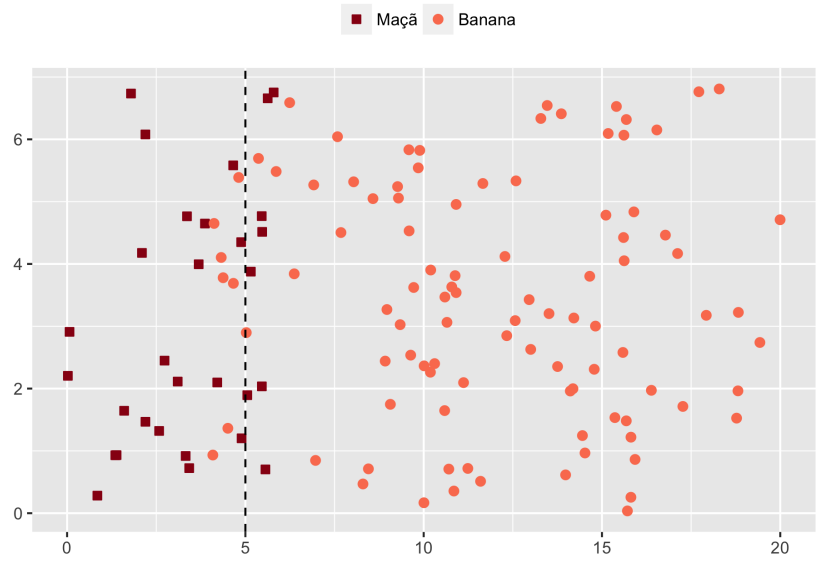

Figura 3.9: Adaptação dos dados para a primeira região de incerteza.

Diante da segunda situação, em que se almeja encontrar a região de incerteza entre banana e laranja, o mesmo procedimento será realizado de modo que, do lado esquerdo, tudo o que não é banana passa a ser caracterizada como tal, e, para o lado direito, tudo o que não for laranja passa a ser considerado como tal.

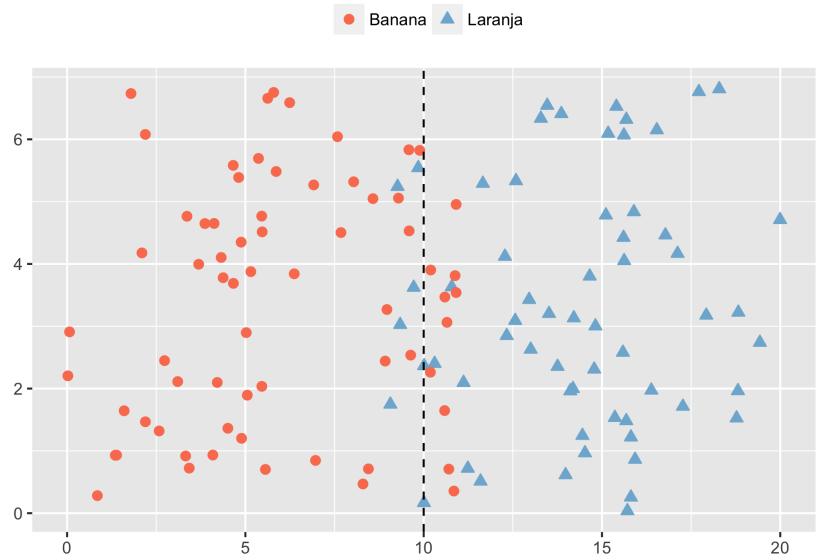

Figura 3.10: Adaptação dos dados para a segunda região de incerteza.

Por fim, para se definir a fronteira entre laranja e maçã-verde, tudo o que está do lado esquerdo passa a ser caracterizado como laranja, e tudo o que se encontra do lado direito será caracterizado como maçã verde.

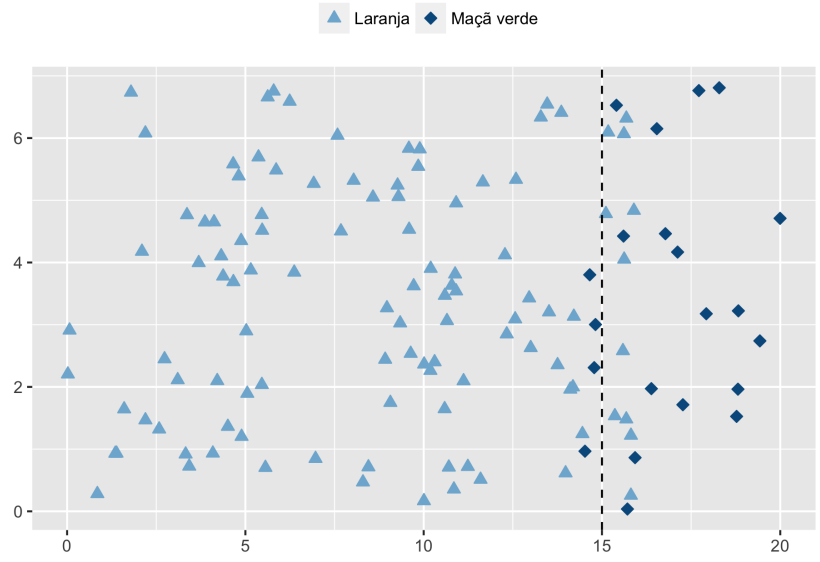

Figura 3.11: Adaptação dos dados para a terceira região de incerteza.

De maneira resumida, tudo o que está de um lado de determinada divisão passa a ser considerado como a categoria mais próxima da barreira de transição, do lado que está sendo considerado. Esse procedimento é realizado porque, deseja-se encontrar as barreiras entre duas categorias adjacentes. De modo que, se forem considerados os demais dados

não pertencentes às categorias de interesse, a melhor divisão dos dados para estas, poderia ser afetada se consideradas as demais. Por fim, após o devido ajustamento dos dados, os índices de interesse são calculados conforme explicado no exemplo com duas categorias.

3.1.2 Regiões de incerteza com base no modelo criado por Rios

A. Barreiras Eliminatórias de reginõe de confusão com base no modelo ajustado

O quarto método para definição das regiões de incerteza é realizado de forma semelhante ao primeiro método descrito. Após a previsão das categorias das amostras de arroz, é possível fazer uma comparação entre a categoria alocada por meio da previsão do modelo e a categoria originalmente escolhida na análise sensorial. No estudo conduzido por Rios, foi possível identificar falhas na classificação das amostras por meio do modelo, em relação às categorias alocadas via sensorial. Sendo assim, há a criação de uma região de confusão de classificação dos dados entre a categorização real e a prevista.

É possível plotar o gráfico de dispersão dos dados com base nos escores obtidos por meio da regressão logística. Dessa maneira, a Figura 3.12 ilustra um exemplo da situação descrita. Os dados estão dispostos com na base classificação real (via sensorial) dos dados, em que os dados representados por pontos abertos, são amostras classificadas erroneamente. Assim dados que possuem a mesma cor representam amostras que foram classificadas na mesma categoria, de acordo com a classificação do modelo.

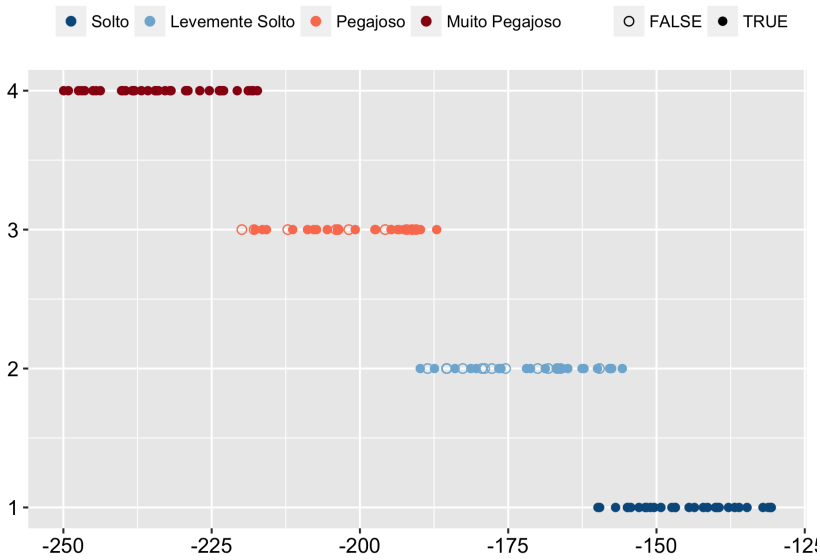

Figura 3.12: Grafico de dispersao de exemplo de classificacao de pegajosidade de acordo com o Escore1 (escore formado pela primeira componente das variaveis de perfil viscoamilografico) [1].

É considerada como incerteza qualquer observação que cair entre os escores que delimitam onde começa e onde termina a confusão de classificação. Ou seja, a partir do ponto em que determinada amostra é classificada segundo a previsão do modelo, e que esta não corresponde à classificação original do arroz via análise sensorial, até o ponto em que esse erro de classificação não é mais identificado amostra será considerada como uma incerteza. Por fim, após a definição das barreiras, com base nesse critério são plotadas as regiões de incerteza no gráfico com as curvas de probabilidades estimadas.

B. Barreiras pré-fixadas

O quarto critério para a definição das regiões de incerteza é estabelecer valores pré-fixados para elas. Neste estudo, o critério para estebelecer essas barreiras requer que a distância entre duas curvas de probabilidade de categorias adjacentes se distanciam igualmente para o lado esquerdo e direito do ponto p_i de encontro dessas curvas de probabilidade de categorias adjacentes. Assim, é determinado um valor fixo para a distância dessas curvas, de modo que admite-se um erro da probabilidade estimada para cima e para baixo do ponto de encontro p_i. Neste estudo, o valor estabelecido para a distância entre as curvas foi 0,2. A Figura 3.13 ilustra a primeira região de incerteza com base no critério estabelecido.

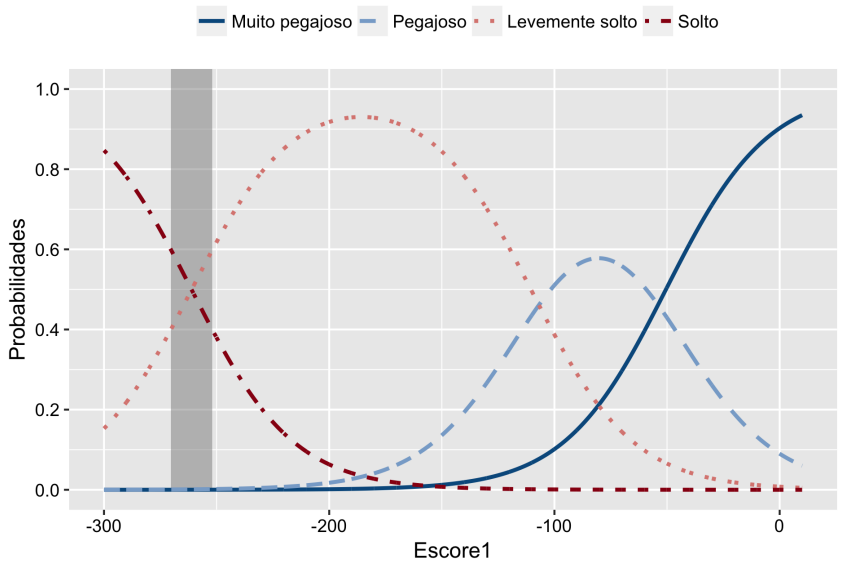

Figura 3.13: Barreira de transição de classificação entre as categorias Solto e Levemente Solto, com base no modelo de Rios [1].

As barreiras de incerteza calculadas são resumidas na Tabela 3.5, com a respectiva notação. As barreiras de Gini e de Entropia são consideradas do tipo (ii), uma vez que ambas são calculadas com base em medidas vindas da Teoria da Informação.

Tabela 3.5: Notações para distintas definições de regiões de incerteza.

Barreira	Notação
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

3.2 Comparação entre tipos de barreiras

A busca da definição das regiões de incerteza é motivada pela melhora da previsão da classificação das amostras de arroz. O estudo busca estabelecer tais regiões de modo que toda amostra que cair dentro dos limites destas seja classificada como incerteza. Todas as possíveis diferentes maneiras de definição das barreiras, que delimitam essas regiões, possuem seus aspectos positivos e negativos. Dessa forma, se faz necessário o uso de medidas que permitam a comparação entre a previsão dada pelo modelo, e o que foi originalmente previsto por meio da análise sensorial, com base no tipo de barreira que foi escolhido. Assim, os erros de classificação após definidas as regiões de incerteza são

obtidos por meio do erro aparente, que são calculados pela Equação 3.16.

$$\text{Erro Aparente} = \frac{\text{Nº total de erros observados fora das regiões de incerteza}}{\text{Nº total de observações fora das regiões de incerteza}} \quad (3.16)$$

Para alguns casos, é possível fazer uma analogia da análise dessas regiões com um intervalo de confiança. Da mesma maneira que nos intervalos de confiança, quando o nivel de significância aumenta, o intervalo fica menor (ou seja, estamos sendo mais criteriosos para a confiança dos dados), para o presente estudo, a medida que a região de incerteza é menor, e o erro de classificação aumenta.

Além disso, quanto maior for a região de incerteza estabelecida, mais amostras de arroz são consideradas como incertezas (não são classificadas), de modo que um custo é atribuído quando ocorre redução dos erros de classificação, de acordo com a barreira estabalecida.

Custo de classificação - segundo tipo de barreira

Com base nos fatos apresentados, foi criada uma medida de custo que avalia o impacto do tipo da região de incerteza escolhida nos resultados da classificação do arroz. O custo atribuído às regiões estabelecidas é dado pela porcentagem de amostras que não foram classificadas, sendo função do avaliador fazer uma decisão e estebelecer qual método deseja empregar com base nos resultados obtidos.

Outra medida que servirá como auxílio para a análise da qualidade das regiões de incerteza, é o número de erros de classificação dentro das regiões estabelecidas. Assim, é feito o cálculo da proporção de erros observados na região (de acordo com o tipo empregado), em relação ao total dos erros originalmente encontrados por meio da comparação da classificação sensorial e a prevista pelo modelo. Essa medida será chamada de "erro proporcional"e calculado por meio da Equação 3.17.

$$\text{Erro Proporcional} = \frac{\text{Nº total de erros observados nas regiões de incerteza}}{\text{Nº total de erros observados originalmente}} \quad (3.17)$$

Esse cálculo é conduzido de modo que se possa fazer a avaliação de quão boa é a definição das regiões de incerteza, em relação a identificação dos erros de classificação segundo os critérios estabelecidos.

Por fim, são encontradas a quantidade de amostras classificadas corretamente dentro das regiões de incerteza definidas e faz-se a análise da proporção de amostras que foram classificadas corretamente (e são consideradas como incerteza), em relação ao total de amostras na região estebelecida. Essa medida será chamada de "Classificação Correta na Região de Incerteza"(com notação Classif. Correta RI) e calculada por meio da Fórmula 3.18.

$$\text{Classif. Correta RI} = \frac{\text{Nº total de amostras classificadas corretamente na região de incerteza}}{\text{Nº total de amostras na região de incerteza}} \tag{3.18}$$

Assim, é feita a análise da qualidade da definição da região de incerteza, considerando quantas amostras classificadas corretamente também estão sendo consideradas como incerteza, segundo o critério estabelecido. Esse cálculo está diretamente relacionado ao custo, uma vez que quanto maior a proporção de amostras classificadas corretamente dentro das regiões, um maior número de amostras não estão sendo classificadas.

3.3 Projeção de barreiras em superfícies - Método Bonferroni

Para o caso da definição da região de incerteza de uma superfície, como ocorre para os dados com duas componentes, se faz necessário o uso da técnica de Bonferroni, para a definição das superfícies de incerteza. Esse método é empregado de maneira análoga ao desenvolvimenteo de intervalos de confiança simultâneos [8].

Em um modelo de regressão, o procedimento que fornece uma família de coeficientes de confiança para estimar β_0 e β_1 é de grande interesse, uma vez que este permite a análise de dois resultados separados em um conjunto de resultados integrados [8]. A construção de intervalos de confiança simultâneos, baseada em uma família específica de coeficientes de confiança para β_0 e β_1 é discutida a seguir, com o uso do procedimento de Bonferroni [8].

A construção da região de incerteza de Bonferroni é feita quando utilizamos um método de predição sensorial com duas variáveis explicativas (Componentes Principais 1 e 2). O intervalo de incerteza é definido a partir dos quatro critérios descritos anteriomente nessa seção para cada eixo (CP1 e CP2), e em seguida é feita a projeção destas. Dessa maneira, a área onde as dois intervalos se cruzam é a região de incerteza de interesse, que corresponde ao elipsoide do exemplo. Essa área é representada no exemplo da Figura

3.14.

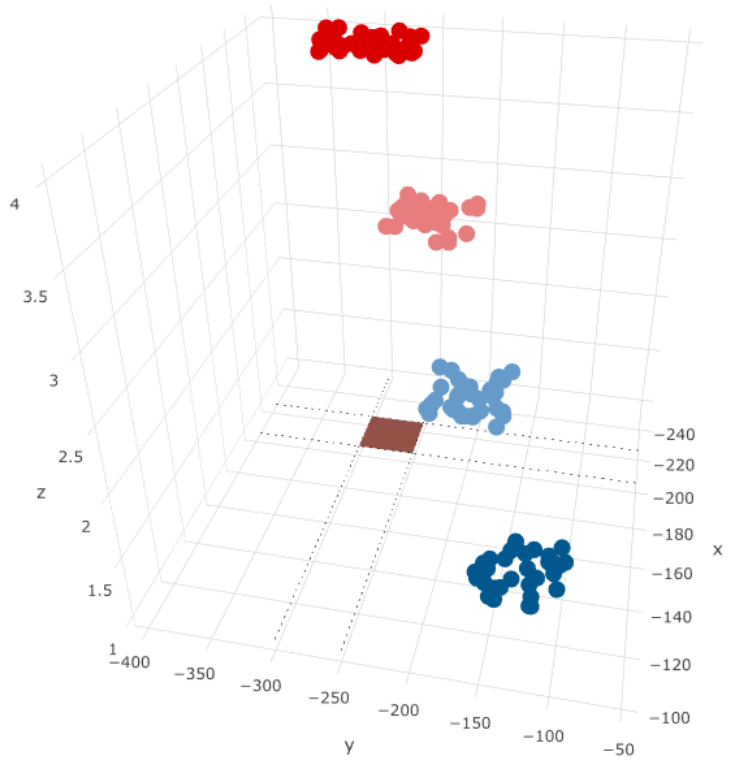

Figura 3.14: Região de incerteza entre categorias Solto e Levemente Solto, para superfície de exemplo de dados, com base nos Escores 1 e 2 previstos pelo modelo, segundo classificação original sensorial da amostra.

A Figura 3.14 apresenta um exemplo de área de incerteza entre as categorias Solto e Levemente Solto. Após a definição das 3 regiões de incerteza, estas são plotadas nos gráficos com as superfícies de probabilidade. Por fim, toda observação que cair nas àreas estabelecidas, serão consideradas como incerteza.

4 Resultados e discussão

4.1 Terras altas

4.1.1 Modelo

O modelo criado para a previsão da classificação sensorial do arroz usa como critério classificador o maior valor entre as probabilidades de uma amostra pertencer a cada categoria. Assim, não há a definição dos limites das regiões de incerteza até então. Com base no modelo, é possível plotar o gráfico de dispersão dos Escores1 obtidos por meio da regressão logística, segundo a classificação prevista por este, ilustrado na Figura 4.1.

Figura 4.1: Gráfico de dispersão dos dados com base nos Escores1 obtidos, segundo classificação prevista pelo modelo para terras altas.

É possível observar que é clara a distinção entre as categorias, a partir do gráfico de dispersão, com base na previsão do modelo. Dessa maneira, a disposição dos dados é idealmente representada no gráfico de dispersão da Figura 4.1. Assim, a transição entre as categorias é de fácil acesso e estabelecida pelas barreiras apresentadas na Tabela 4.1 e ilustradas na Figura 4.2.

Tabela 4.1: Limites de barreiras de transição segundo Escores1 preditos, para terras altas.

Transição	Limite
S - LS	x
LS - P	-107
P - MP	-54

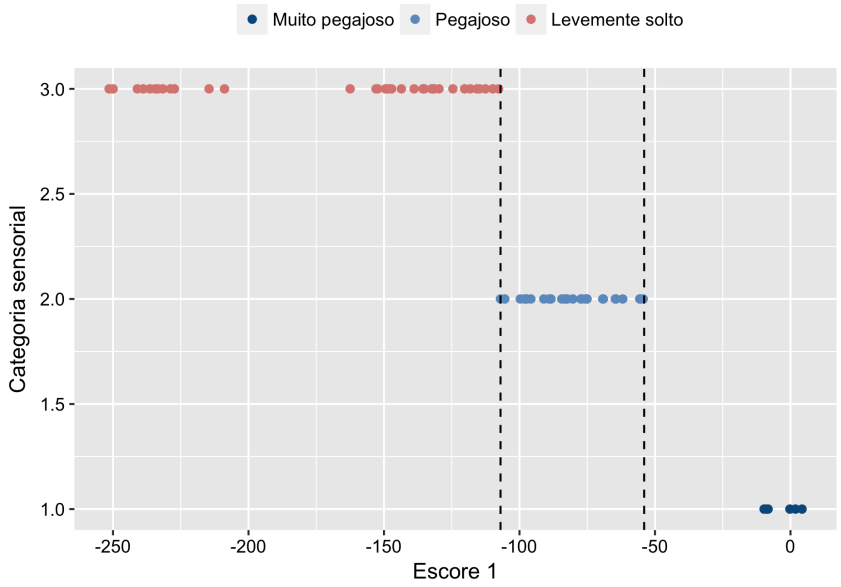

Figura 4.2: Barreiras de transição de categorias com base nos Escores1 obtidos, segundo classificação prevista pelo modelo para terras altas.

Assim, ao plotar as barreiras apresentadas na Tabela 4.1, junto ao gráfico com as curvas de probabilidade preditas, esses limites se encontram exatamente nos pontos de transição p_i das curvas de categorias adjacentes, ilustrados na Figura 4.3.

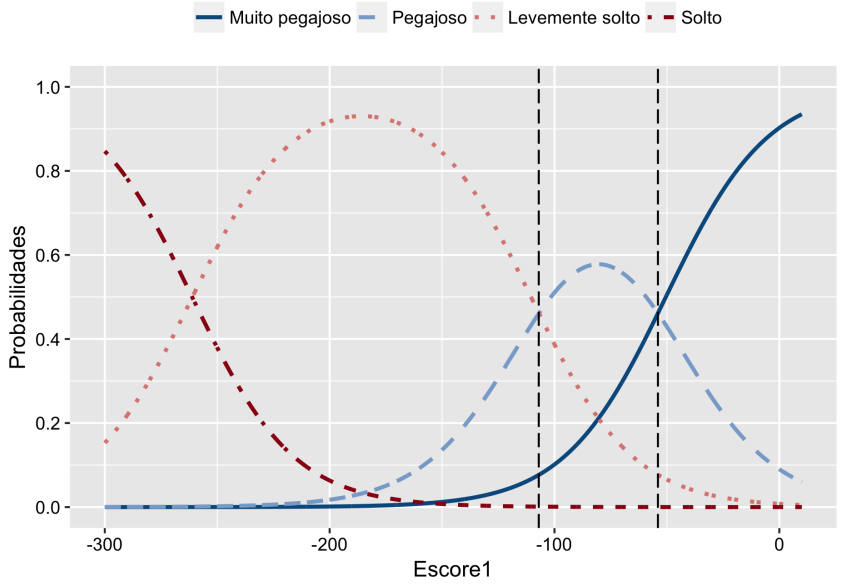

Figura 4.3: Barreiras de transição de curvas de probabilidades adjacentes.

Para o caso da transição entre as categorias Solto (S) e Levemente Solto (LS), não houve nenhuma amostra prevista na categoria S, de modo que não há como traçar a barreira de transição entre essas categorias. Esse fato pode ser explicado pela baixa

frequência de dados originalmente classificadas como Soltos, em que apenas três amostras foram categorizadas como tal.

A partir da previsão do modelo e das classificações sensoriais conhecidas do arroz, é possível fazer a análise do erro de classificação. A Tabela 4.2 apresenta o estudo de validação cruzada para esse caso [1].

Tabela 4.2: Classificação sensorial de pegajosidade do arroz versus a classificação prevista, segunod modelo para terras altas.

		Classe prevista			
	Categoria	MP	P	LS	S
Classe real	MP	6	7	0	0
	P	0	18	3	0
	LS	0	4	31	0
	S	0	0	3	0

O erro de classificação obtido é de 23,61%. O custo de 0% é atribuído a esse critério, de modo que todas as amostras são classificadas.

4.1.2 Barreiras Eliminatórias com base na dispersão real

Para esse caso, os limites das barreiras de incerteza para cada região são definidos a partir do ponto onde a sobreposição de dados de duas categorias adjacentes começa, até onde ela não está mais presente. A partir do gráfico de dispersão dos dados de terras altas, segundo a classificação obtida via sensorial, representada na Figura 4.4, é possível observar onde essa sobreposição ocorre.

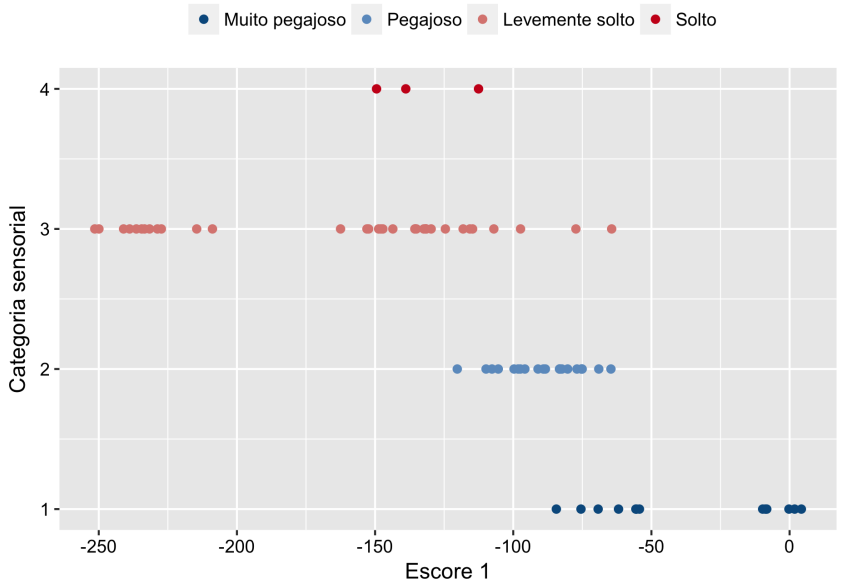

Figura 4.4: Gráfico de dispersão dos dados com base nos Escores1 obtidos para os dados de terras altas, segundo classificação via sensorial.

Dessa maneira, os limites dessas regiões são obtidos, segundo tal critério, para as três transições entre categorias adjacentes, apresentados na Tabela 4.3. A primeira região de incerteza ocorre na transição das categorias Solto (S) e Levemente Solto (LS), a segunda na transição entre Levemente Solto (LS) e Pegajso (P) e, por fim, a terceira região na transição entre Pegajoso (P) e Muito Pegajoso (MP).

Tabela 4.3: Limites das regiões de incerteza baseadas nos Escores1, segundo método BER.

Região de incerteza	Limite inferior	Limite superior	Amplitude
S - LS (1)	-149,4753	-112,5598	36,9155
LS - P (2)	-120,2610	-64,6551	55,6059
P -MP (3)	-84,4076	-64,6551	19,7525

Com base nos limites das regiões, representados pelos Escores1 previstos na regressão logística, o gráfico com as barreiras é representado pela Figura 4.5.

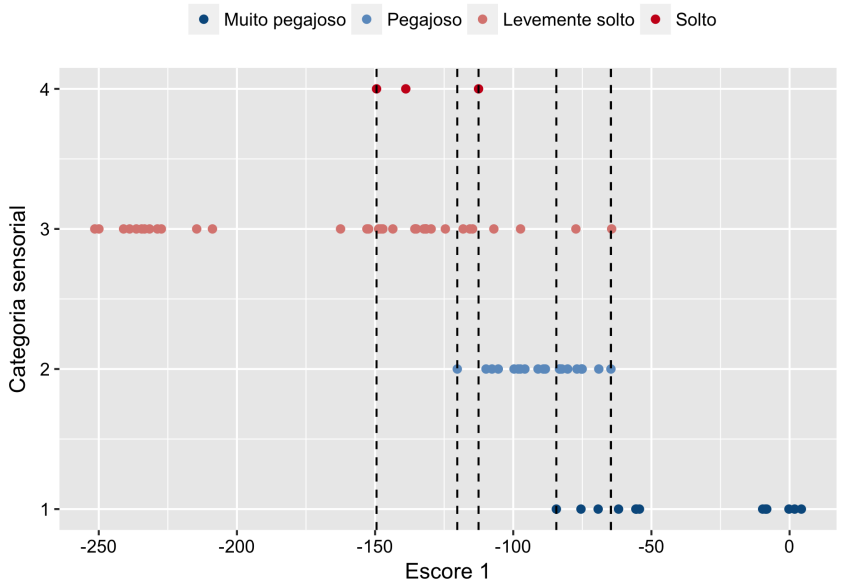

Figura 4.5: Limites das regiões de incerteza baseadas nos Escores1, segundo método BER.

Assim, as regiões de incerteza podem ser encontradas, conforme ilustrado na Figura 4.6 e 4.7.

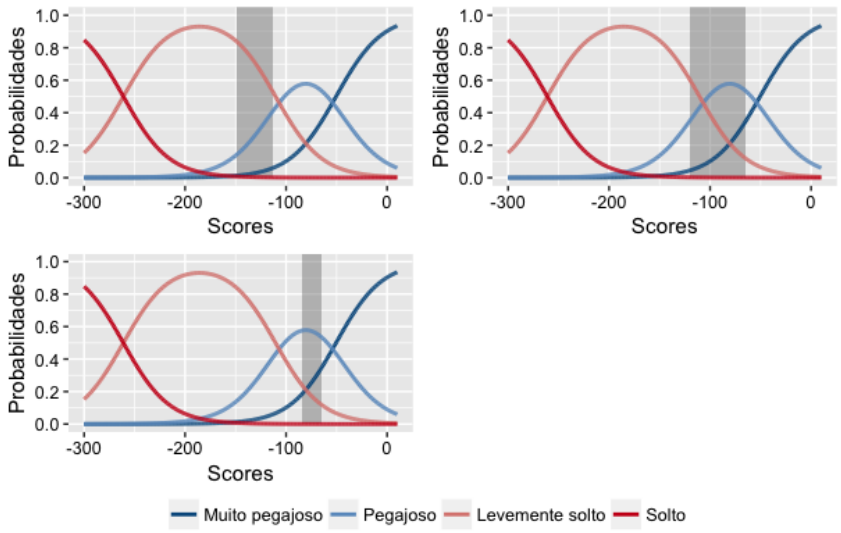

Figura 4.6: Regiões de incerteza, segundo método BER.

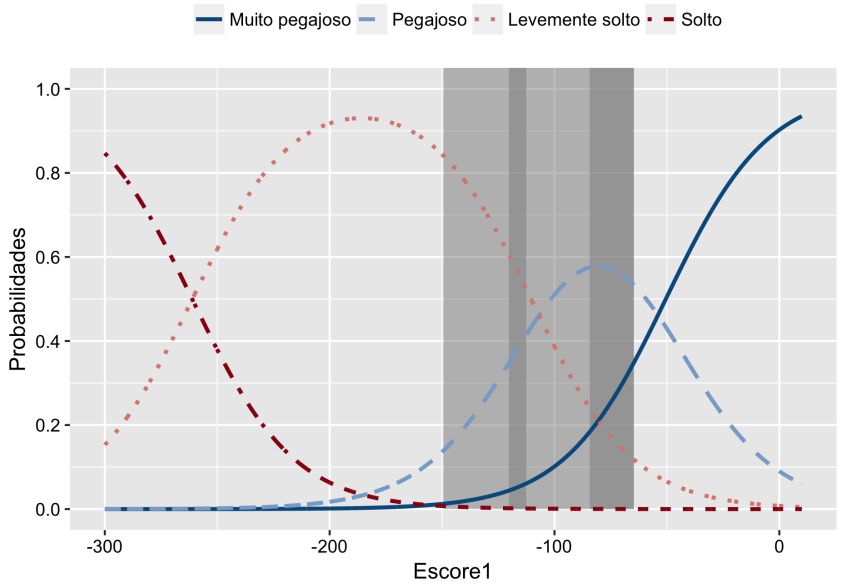

Figura 4.7: Regiões de incerteza, segundo método BER.

Por meio das figuras apresentadas, é possível observar que a sobreposição dos dados é muito presente. Entretanto, ainda é possível fazer uma distinção na transição das categorias, com base no gráfico de dispersão representado pela Figura 4.4. Algumas regiões de incerteza ocorrem dentro dos limites de outras. Esse acontecimento pode ser exemplificado pelas amostras classificadas como Soltas, em que as três observações estão extremamente dentro da área da sua categoria adjacente — Levemente Solto. Isso ocorre pelo fato dos dados estarem muito sobrepostos, o que pode ser uma indicação de que o avaliador está errando muito na classificação da amostra de arroz, ou de que o modelo não está fazendo a previsão dos dados da melhor maneira.

Pode-se observar, por meio da Figura 4.6, que as regiões de incerteza não se encontram exatamente em torno dos pontos p_i de transição das curvas de probabilidade, em que essa seria a situação ideal, conforme exemplificado anteriormente na Figura 4.1 (seção 4.1.1). É possível observar que para a primeira região de incerteza, entre as categorias S e LS essa região se encontra excesivamente distante do ponto de transição das curvas entre essas duas categorias, pelo fato do modelo não ter previsto nenhuma amostra como Solta. Para a segunda região, na transição entre LS e P, a região engloba o ponto de transição p_2; entretanto, ainda ocorrem erros distantes desse ponto. Por fim, é possível observar que a terceira região se encontra perto do ponto p_3 de transição entre as categorias P e MP. Entretanto, ainda há a ocorrência de erros distantes desse ponto.

Assim, a Figura 4.6 serve como auxílio para a análise de onde as regiões de incer-

teza se encontram nas curvas de probabilidade, e o quão próximas as regiões estabelecidas estão do ponto p_i de transição, com base no critério utilizado em questão. Situações em que as regiões que estão muito distantes do ponto p_i de transição entre as categorias também podem ser um indicativo de que o modelo não está prevendo da melhor maneira as amostras, ou de que a classificação sensorial do arroz ainda pode ser melhorada.

De acordo com a análise da curva ROC (Figura 4.8) para os dados de terras altas [1], a adequação da previsão do modelo para a categoria Solta é fraca, de modo que as observações para esta categoria se encontram muito inseridas nos limites da sua categoria adjacente (LS). As amostras classificadas como MP, de acordo com a curva ROC, são as observações em que o modelo executa melhores predições Assim, a transição entre as categorias MP e P são as que menos se sobrepoõem no gráfico de dispersão.

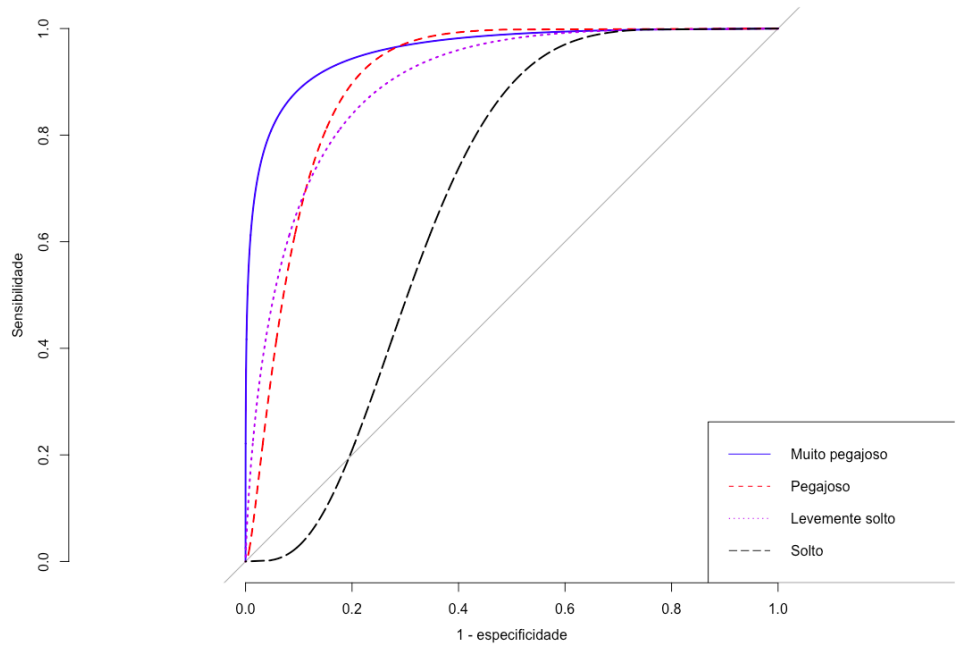

Figura 4.8: Curva de classificação ROC da avaliação sensorial de pegajosidade para o ano de 2014, prevista por meio do modelo de regressão logística, utilizando componentes principais de arroz de Terras Altas.
Nota: Figura de Rios [1]. Autorizado pelo autor.

Com relação a transição entre as categorias Pegajoso e Levemente Solto , pode-se observar que a primeira está totalmente inserida na região da segunda, entretanto, isso ocorre de maneira menos problemática do que para a categoria Solto. Nesse caso, pode-se notar que há a presença de alguns pontos discrepantes para a classe LS, o que explica a total inserção da categoria P em LS. A presença dos pontos discrepantes para a classe LS pode ser explicada pelo fato de que a previsão do modelo para essa categoria é boa.

Entretanto, não é a melhor entre elas, como no caso da categoria MP [1].

Por fim, a grande sobreposição dos dados faz com que os limites das regiões de incerteza fiquem muito amplos. Dessa maneira, por meio da Figura 4.7 é possível notar que não há uma distinção clara entre as regiões de incerteza.

Após encontradas as regiões de interesse, é possível analisar a redução do erro aparente após a retirada das amostras consideradas como incerteza. Dessa maneira, por meio da Tabela 4.4, o erro de classificação pode ser determinado.

Tabela 4.4: Classificação sensorial de pegajosidade do arroz versus a classificação prevista, segundo método BER.

		Classe prevista			
	Categoria	MP	P	LS	S
Classe real	MP	6	4	0	0
	P	0	0	0	0
	LS	0	1	15	0
	S	0	0	0	0

A Tabela 4.5 apresentada a seguir, mostra os resultados do erro aparente encontrado por meio da Tabela 4.4 e o custo associadao às barreiras de incerteza. Além disso, a tabela apresenta a proporção de erros em relação ao total de discordâncias originalmente observadas, e a proporção de amostras classificadas corretamente entre os limites dessa região.

Tabela 4.5: Medidas auxiliares para avalição das regiões de incerteza.

Erro de class.	Custo	Erro prop.	Class. correta
19,23%	63,89%	70,59%	73,91%

Assim, toda amostra que cair na região de incerteza definida, não será alocada em nenhuma categoria. Para este caso, o erro de classificacação reduziu de 23%, para aproximadamente 19,23%. O custo associado à redução da incerteza do modelo resultou na não classificação de 63,89% das amostras, ou seja, elas estão dentro dos limites das regiões, e, portanto, consideradas como incerteza.

Observa-se que as regiões definidas com base na dispersão real dos dados, implica na retirada de 70,59% dos erros de classificação. Assim, pode-se concluir que uma grande parcela dos erros de classificação se encontram em áreas onde ocorrem a sobreposição de dados entre categorias adjacentes. Entretanto, na região estabelecida, 73,91% dos

dados são categorias classificada corretamente. Dessa maneira, o erro de classificação não reduziu muito, de 23% para 19,23%, pelo fato de que muitas amostras classificadas corretamente foram desconsideradas, aumentando a proporção de erros em comparação ao total de observações (que foi muito reduzido).

Além disso, é possíver fazer a análise de qual das três regiões está associada ao maior erro de classificação do modelo. A Tabela 4.6 apresenta o erro de classificação e o custo considerando a retirada uma-a-uma das regiões de incerteza estabelecidas.

Tabela 4.6: Comparação entre resultados com regiões retiradas uma-a-uma, segundo método BER.

Região retirada	Erro de classif.	Custo	Erro prop.	Classif. correta RI
Região 1 (S -LS)	25%	27,78%	23,53%	80%
Região 2 (LS-P)	17,07%	43,06%	58,82%	67,74%
Regiao 3 (P-MP)	22,41%	19,44%	23,53%	71,43%

Por meio da Tabela 4.6, observa-se que quando considerada apenas a segunda região, o erro pode ser reduzido para 17,07%. O custo da classificação é de 43,06%, em que a amplitude dessa região é a maior entre as três (Tabela 4.3), fazendo com que mais observações sejam consideradas como incerteza. Isso é confirmado pelo erro proporcional encontrado, no valor de 58,82%, o que mostra que essa região contêm aproximadamente a metade dos erros totais de classificação.

Após a análise da definição das regiões, foi possível observar um erro de classificação de 19,23% e um custo de 63,89% das amostras não classificadas (tipo (i)). Ao retirar apenas a região que traz maior impacto no erro de classificação, neste caso a segunda região, pode-se observar que o erro e o custo da classificação foram menores do que quando considerada as três regiões, resultam nos valores de 17,07% e 43,06%, respectivamente (tipo (ii)). A comparação entre essas duas situações estão aparesentadas na Tabela 4.7.

Tabela 4.7: Comparação entre resultados por tipo de critério, segundo método BER.

Regiões de incerteza	Erro de classif.	Custo
Tipo (i)	19,23%	63,89%
Tipo (ii)	17,07%	43,06%

Pode-se concluir que a retirada apenas da segunda região de incerteza que é a mais vantajosa, uma vez que o erro e o custo são menores para essa situação. Assim, as análises realizadas daqui em diante, com base neste critério, serão baseadas na situação do tipo

(ii).

4.1.3 Medidas de categorização de dados

I. Gini - BGI

Após conduzida todas as possíveis divisões dos dados, utilizando o Índice de Gini foi possível obter as duas melhores divisões, ou seja, os dois índices com valores mais próximos de zero, apresentados na Tabela 4.8. Os valores do Gini 1 e Gini 2 representam os dois menores resultados encontrados, respectivamente.

Tabela 4.8: Índices de Gini encontrados para os dois menores resultados por região de incerteza, segundo método BGI.

Região de incerteza	Gini 1	Gini 2
S - LS (1)	0,0761	0,0763
LS - P (2)	0,1260	0,1475
P -MP (3)	0,0893	0,1040

As melhores divisões encontradas são estabelecidas. Os limites das regiões de incerteza são representados pelos Escores1, preditos pelo modelo, onde ocorrem os dois menores valores do índice, para cada transição de categoria, apresentadas na Tabela 4.9.

Tabela 4.9: Limites das regiões de incerteza baseadas nos Escores1, segundo método BGI.

Região de incerteza	Limite inferior	Limite superior	Amplitude
S - LS (1)	-109,8510	-107,7004	2,1506
LS - P (2)	-112,5598	-109,8510	2,7088
P -MP (3)	-61,8988	-55,6344	6,2644

Dessa maneira, as barreiras das regiões são ilustradas na Figura 4.9.

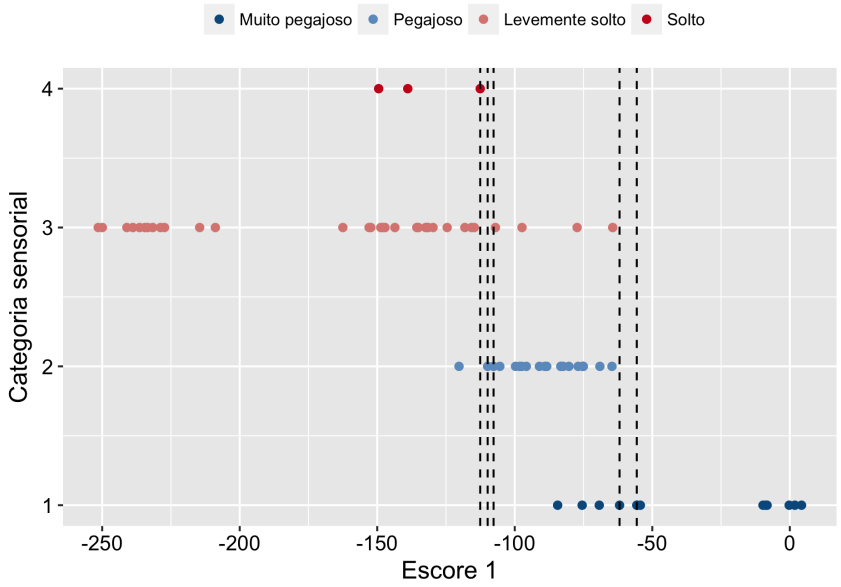

Figura 4.9: Regiões de incerteza, segundo método BGI.

E assim, são plotadas as regiões de incerteza, apresentadas nas Figuras 4.10 e 4.11.

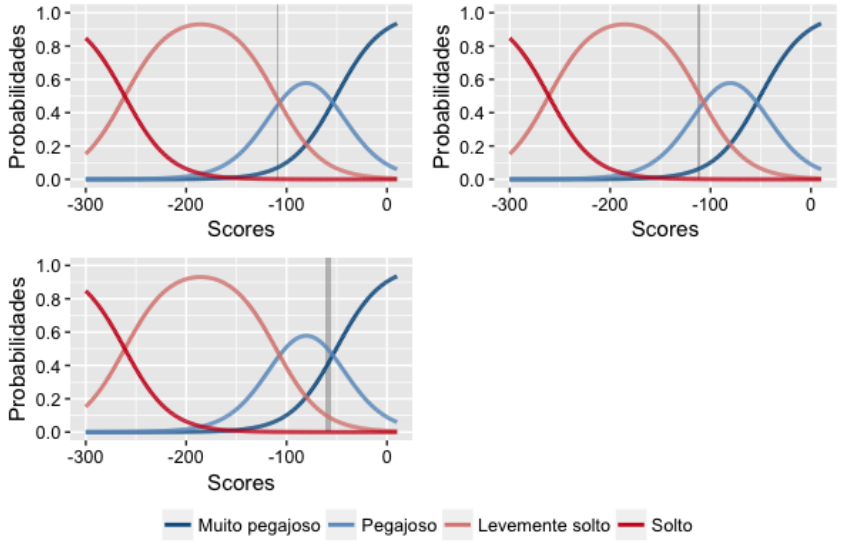

Figura 4.10: Regiões de incerteza segundo método BGI.

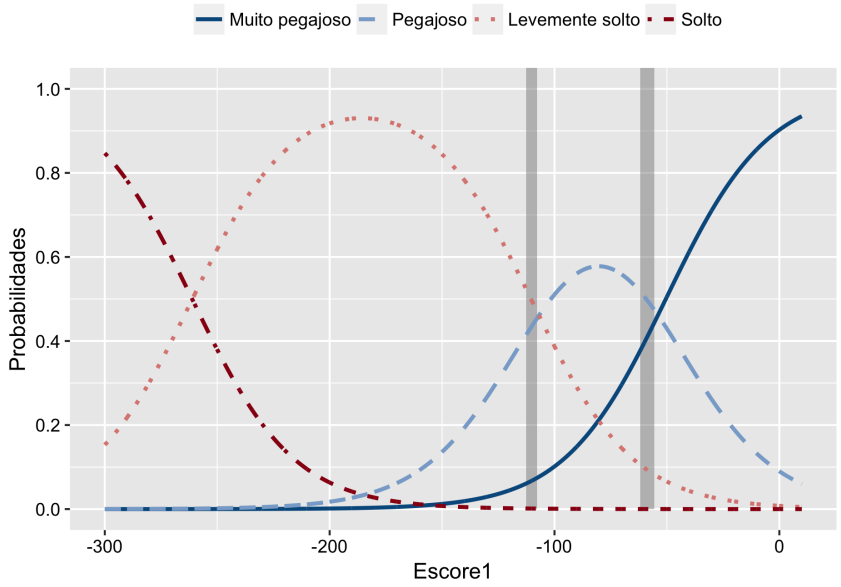

Figura 4.11: Regiões de incerteza segundo método BGI.

A distinção entre as regiões de incerteza (1) e (2) não é clara. A partir da Tabela 4.9 é possível observar que o limite em que uma das regiões termina é onde o limite da outra começa. Nesse caso, a segunda (LS - P) e a terceira (P - MP) regiões de incerteza estão muito próximas dos pontos p_2 e p_3 de transição entre as curvas de categorias adjacentes, respectivamente. Em seguida, a análise do erro aparente é obtida a partir da Tabela 4.10.

Tabela 4.10: Classificação sensorial de pegajosidade do arroz versus a classificação prevista, segundo método BGI.

		Classe prevista			
	Categoria	MP	P	LS	S
Classe real	MP	6	5	0	0
	P	0	18	1	0
	LS	0	4	31	0
	S	0	0	2	0

O erro de classificação nesse caso é reduzido para 17,91%. Assim, um custo de 6,94% das amostras classificadas como incerteza é associado a esse critério.

I. Entropia - BEN

De maneira semelhante, o Critério de Entropia é computado para todas as possíveis divisões do gráfico de dispersão, em que os dois menores valores para cada região são apresentados na Tabela 4.11.

Tabela 4.11: Índices de Entropia para os dois menores valores encontrados, segundo região de incerteza, segundo método BEN.

Região de incerteza	Entropia 1	Entropia 2
S - LS (1)	0,2309	0,2325
LS - P (2)	0,3449	0,3868
P -MP (3)	0,3068	0,3200

Por meio destes, são encontrados os valores dos limites das barreiras, com base nos Escores1, em que as melhores divisões ocorreram, apresentadas na Tabela 4.12.

Tabela 4.12: Limites das regiões de incerteza baseadas nos Escores1, segundo método BEN.

Região de incerteza	Limite inferior	Limite superior	Amplitude
S - LS (1)	-135,6125	-135,4348	0,1777
LS - P (2)	-112,5598	-109,8510	2,7088
P -MP (3)	-69,2740	-64,3876	4,8864

E assim, são encontradas as barreiras e as regiões de incerteza, ilustradas na Figura 4.12, 4.13 e 4.14.

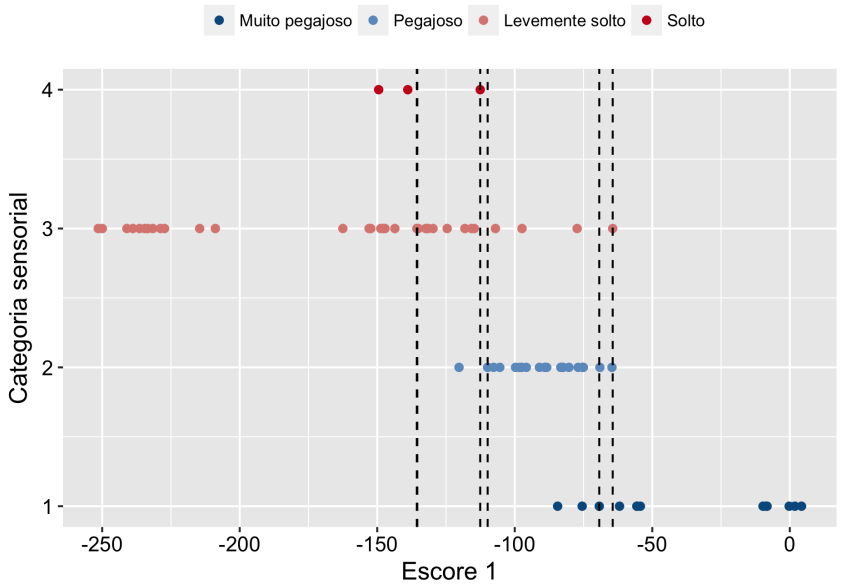

Figura 4.12: Limites das regiões incerteza segundo método BEN.

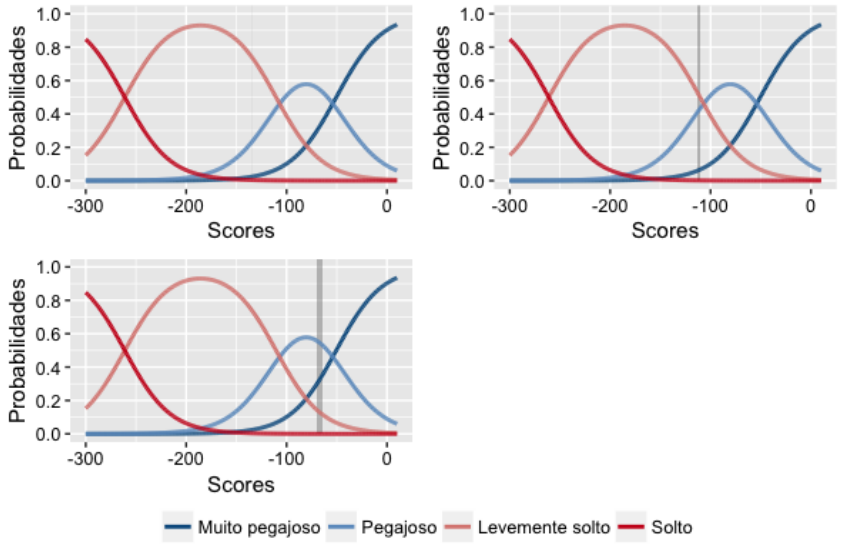

Figura 4.13: Regiões de incerteza da baseadas nos Escores1, segundo método BEN.

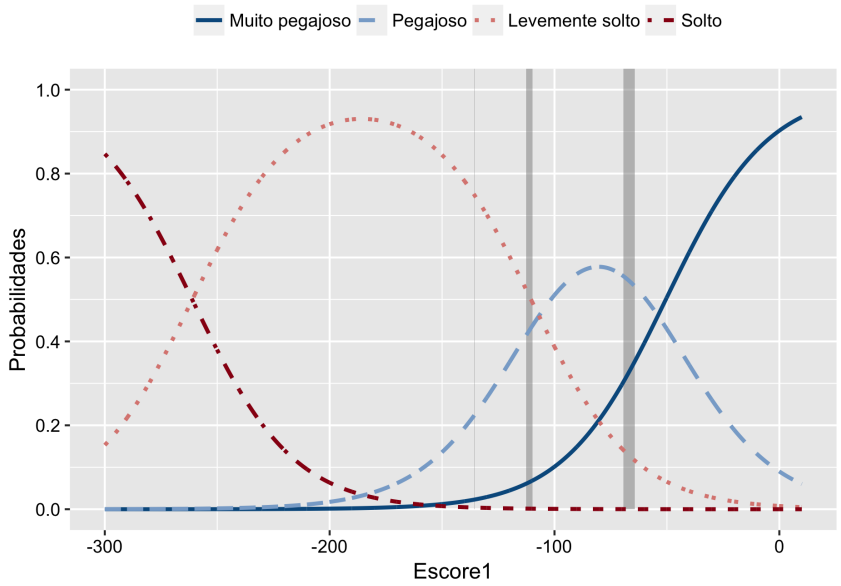

Figura 4.14: Regiões de incerteza segundo método BEN.

O erro aparente para este caso é conduzido por meio da Tabela 4.13.

Tabela 4.13: Classificação sensorial de pegajosidade do arroz versus a classificação prevista, segundo método BEN.

		Classe prevista			
	Categoria	MP	P	LS	S
Classe real	MP	6	6	0	0
	P	0	16	2	0
	LS	0	3	29	0
	S	0	0	2	0

O erro de classificação utilizando o Critério de Entropia é de 20,31%. O custo para a redução do erro de classificação se deu em 11,11% das amostras consideradas como incerteza.

Após a análise de ambos os critério vindos da Teoria da informação utilizados como métodos de classificação, é possível fazer uma comparação entre os dois, apresentada na Tabela 4.14.

Tabela 4.14: Comparação entre resultados segundo método BGI e BEN.

Critério	Erro	Custo	Erro prop.	Classif. correta RI
Gini	17,91%	6,94 %	29,41%	0%
Entropia	20,31%.	11,11 %	23,53%	50%

É possível concluir que, para todas as medidas apresentadas, o índice de Gini é o mais vantajoso entre os dois, tendo como erro e custo de classificação menores do que o de Entropia. O erro proporcional é menor para o índice de Gini, entretanto esse método não considera nenhuma observação classificada corretamente, resultando em um erro menor do que o de Entropia. Sendo assim, daqui em diante o índice de Gini será utilizado para a definição das regiões de incerteza, com base em critérios de classificação de dados.

Assim, faz-se a análise da redução do erro de classificação com a retirada uma-a-uma das três regiões estabelecidas.

Tabela 4.15: Comparação entre resultados com regiões retiradas uma-a-uma, segundo método BGI.

Região retirada	Erro de classif.	Custo
Região 1 (S -LS)	21,43%	2,78%
Região 2 (LS-P)	21,43%	2,78%
Regiao 3 (P-MP)	21,43%	2,78%

Para este caso, cada uma das três regiões contêm apenas duas observações. Sendo assim, o erro e o custo para cada uma delas são os mesmos. Assim, é mais vantajosa a retiradas de todas as áreas estabelecidas, uma vez que o erro é menor e o custo na classificação não é muito grande, no valor de 6,94%.

4.1.4 Barreiras Eliminatórias com base no modelo ajustado

As regiões de incerteza definidas nesta seção utilizam como critério todas as observações em que a classificação sensorial difere da classificação do modelo, segundo os

Escores1 preditos. Esses limites são apresentados na Tabela 4.16.

Tabela 4.16: Limites das regiões de incertezada baseadas nos Escores1, segundo método BEM.

Região de incerteza	Limite inferior	Limite superior	Amplitude
S - LS (1)	-149,4753	-112,5598	36,9155
LS - P (2)	-120,2610	-64,3876	55,8734
P -MP (3)	-84,4076	-54,3335	30,0741

A representação gráfica desta situação é ilustrada na Figura 4.15, em que observações da mesma cor foram classificadas na mesma categoria segundo o modelo. As observações que estão repesentadas por uma ponto vazio, são as observações que o modelo classificou erroneamente, e estão dispostas na categoria original da análise sensorial.

Figura 4.15: Gráfico de dispersão dos dados segundo classificação método BEM.

As barreiras de cada região e as regiões de incerteza nas Figuras 4.16, 4.17 e 4.18.

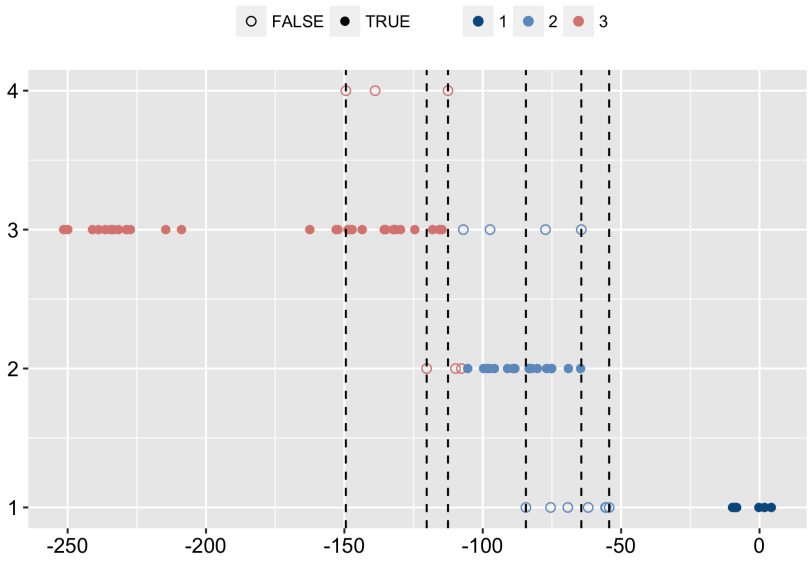

Figura 4.16: Limites das barreiras de incerteza, segundo método BEM.

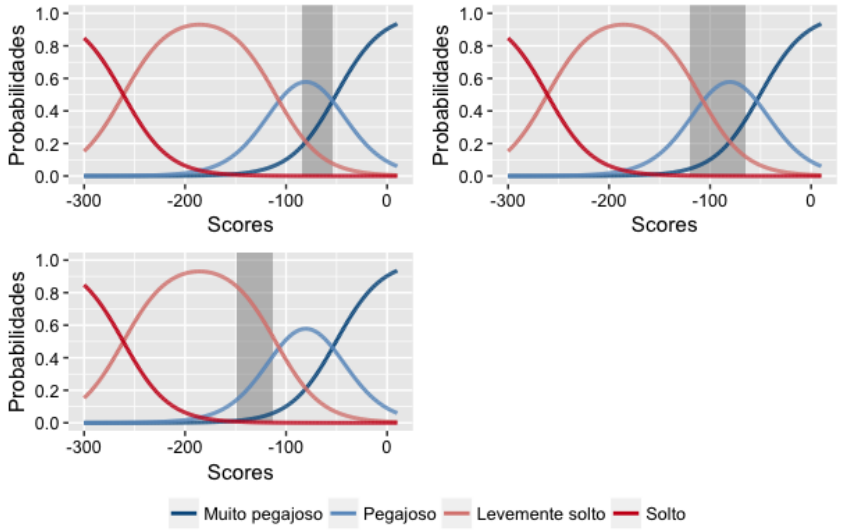

Figura 4.17: Regiões de incerteza segundo métido BEM.

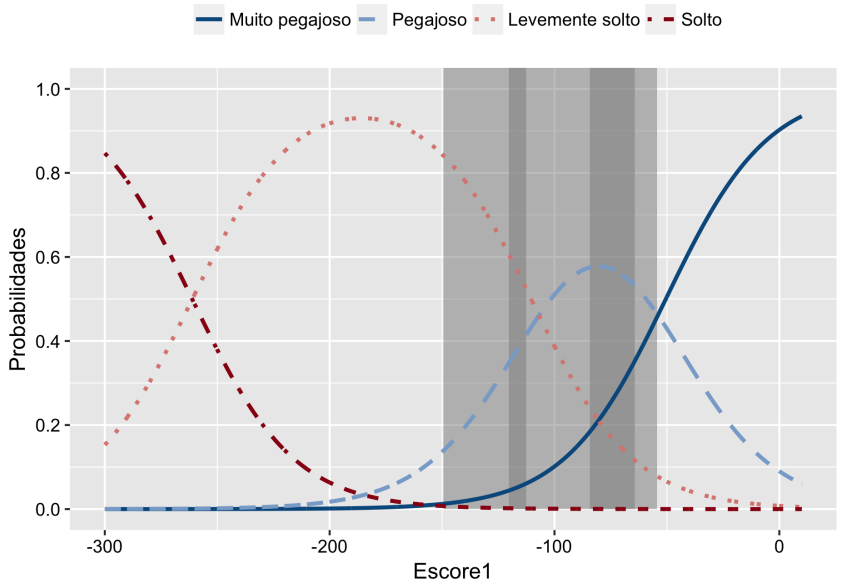

Figura 4.18: Regiões de incerteza segundo método BEM.

A Tabela 4.17 apresenta as informações necessárias para o cálculo do erro aparente dessa situação.

Tabela 4.17: Classificação sensorial de pegajosidade do arroz versus a classificação prevista, segundo método BEM.

		Classe prevista			
	Categoria	MP	P	LS	S
Classe real	MP	6	0	0	0
	P	0	0	0	0
	LS	0	0	15	0
	S	0	0	0	0

Tabela 4.18: Medidas auxiliares para avaliação das regiões de incerteza, segundo método BEM.

Erro de class.	Custo	Erro prop.	Class. correta
0%	70,83%	100%	66,67%

O erro de classificação para este caso é de 0%, pois todas as amostras classificadas erroneamente são retiradas. O custo dessa redução do erro é em 70,83% das amostras não classificadas. Muitas amostras que foram classificadas corretamente, estão entre os limites onde as discordâncias ocorrem, resultando em muitos dados não classificados. Uma vez que todos os erros de classificação foram considerados para essa região de incerteza, 100% dos erros se encontram nessa região, e 66,67% dos dados estão classificados corretamente dentro dos limites das regiões e foram considerados como incertezas.

De maneira similar ao que foi realizado nas regiões de incerteza com base na dispersão real (via sensorial) dos dados, é possível realizar a análise de qual das regiões influencia mais no erro de classificação. Assim, a Tabela 4.19 apresenta o erro e o custo de classificação com a retirada uma-a-uma das regiões de incerteza, assim como as demais medidas de interesse.

Tabela 4.19: Comparação entre resultados com regiões retiradas uma-a-uma segundo método BEM.

Região retirada	Erro de classif.	Custo	Erro prop.	Classf. correta
Região 1 (S -LS)	25%	27,78%	23,53%	80%
Região 2 (LS-P)	15%	44,44%	64,71%	65,52%
Regiao 3 (P-MP)	15,09%	26,39%	52,94%	52,63%

Nesse caso, a região em que ocorre o maior número de erros é a segunda. O erro foi reduzido de 23% para 15%, e com um custo de 44,44%. Entretanto, pode-se observar que para a retiarada da terceira região, o custo é mais baixo (no valor de 26,39%), e o erro é muito próximo ao da segunda região (no valor de 15,09%). Os resultados para este caso são muito semelhantes aos observados no critério embasado no gráfico de dispersão real dos dados (seção 4.1.2). Apesar da retirada de apenas a terceira região ser mais vantajosa neste caso, para fins de comparação são mantidas as três regiões de incerteza que retiram todas as inconsistências de classificação, de modo que se possa obter erro zero na classificação.

4.1.5 Barreiras pré fixadas

Para o caso das regiões de incerteza em que se admite uma margem de erro para cima e para baixo do ponto de encontro p_i entre as curvas de probabilidades preditas, as barreiras de cada região são descritas por meio da Tabela 4.20 e ilustradas pela Figura 4.19.

Tabela 4.20: Limites das regiões de incertezada baseadas nos Escores1, segundo método BPF.

Região de incerteza	Limite inferior	Limite superior	Amplitude
S - LS (1)	-270,32	-251,87	18,45
LS - P (2)	-117,2	-95,20	22,22
P -MP (3)	-65,66	-43,64	22,27

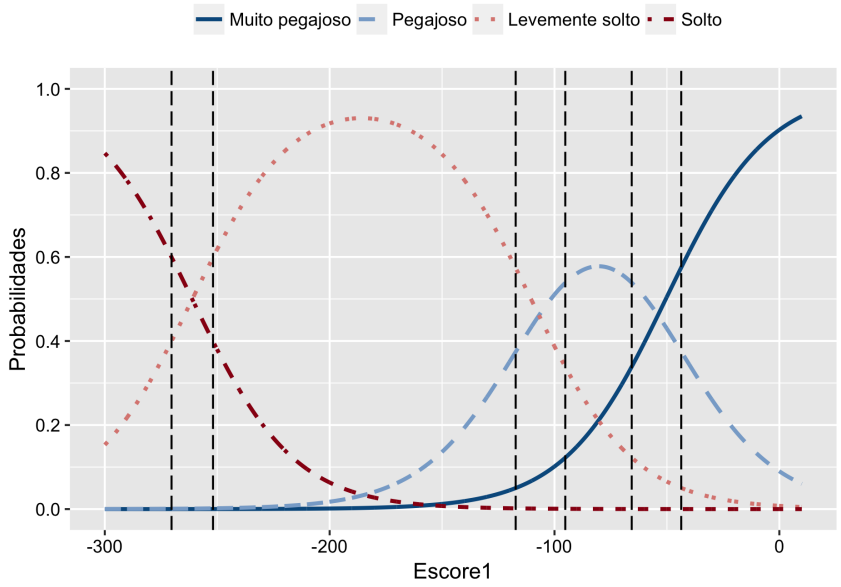

Figura 4.19: Limites das regiões de incerteza segundo método BPF.

As regiões de incerteza são encontra das e ilustradas na Figura 4.20.

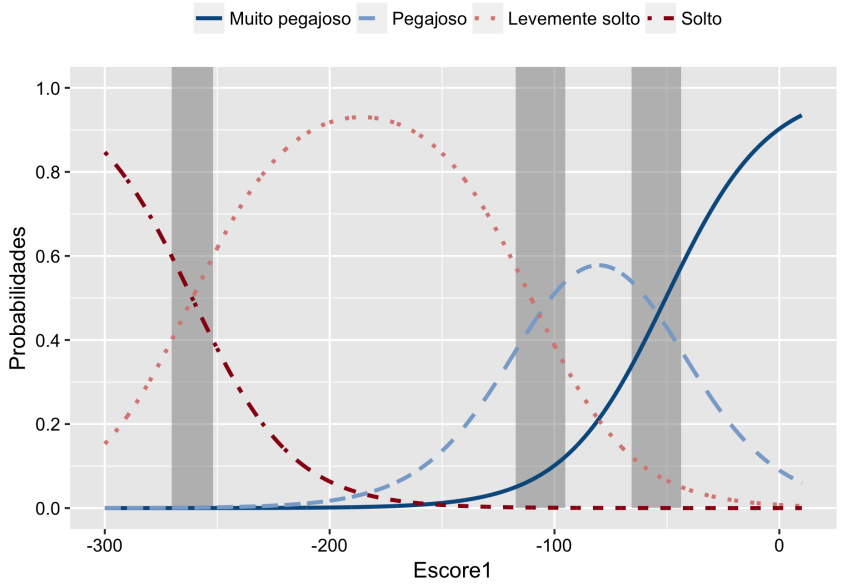

Figura 4.20: Regiões de incerteza segundo método BPF.

A análise do erro aparente é conduzido por meio da Tabela 4.21.

Tabela 4.21: Classificação sensorial de pegajosidade do arroz versus a classificação prevista, segundo método BPF.

		Classe prevista			
Categoria	MP	P	LS	S	
Classe real	MP	6	3	0	0
	P	0	12	1	0
	LS	0	1	29	0
	S	0	0	2	0

Tabela 4.22: Medidas auxiliares para avaliação das regiões de incerteza, segundo método BPF.

Erro	Custo	Erro prop.	Classif. correta RI
12,96%	25%	58,82%	44,44%

A taxa de erro de classificação é dada por 12,96% das amostras classificadas erroneamente, utilizando este critério como definição das regiões de incertezas. O custo para a redução do erro resultou em 25% das observações não alocadas em classe nenhuma, ou seja, são consideradas incertezas. Neste caso, por meio dos Escores1 obtidos na regressão logística, notou-se que não houve a presença de observações dentro dos limites da primeira região, pois nenhum dado foi classificado como Solto pelo modelo. Assim, no cálculo para a análise do erro e custo de classificação não foi considerada a primeira região de incerteza (S - LS).

A análise do impacto do erro em cada uma das regiões estabalecidas, apresentadas na Tabela 4.23, é realizadas a seguir.

Tabela 4.23: Comparação entre resultados com regiões retiradas uma-a-uma, segundo método BPF.

Região retirada	Erro de classif.	Custo	Erro prop.	Classif. correta RI
Região 1 (S -LS)	x	x	x	x
Região 2 (LS-P)	20%	16,67%	29,41%	58,33%
Regiao 3 (P-MP)	18,18%	8,33%	29,41%	16,67%

Por meio dos Escores1 obtidos na regressão logística, notou-se que não houve a presença de observações dentro dos limites da primeira região, pois nenhum dado foi classificado como Solto pelo modelo. A análise da Tabela 4.23 indica que a terceira região de incerteza (P - MP) tem o maior impacto no erro de classificação, em que o menor erro foi observado quando considerada apenas essa região. Além disso, pode-se observar que o custo é baixo nessa situação, tendo o valor de 8,33%.

Nesse caso, é mais vantajosa a retirada da três regiões, uma vez que o erro é reduzido em aproximadamente a metade do erro inicial (23%), e o custo não é muito alto, sendo igual a 25%. Apesar do custo de se retirar apenas a terceira região estabelecida ser menor, nota-se que a redução do erro inicial não é muito grande, sendo reduzido para 18%. Assim, para as análises realizadas no restante do relatório, para este critério, serão consideradas a segunda e a terceira região de incerteza estabelecidas, uma vez que não foi observada nenhuma amostra classificada entre os limites da primeira região.

Vale ressaltar que esse é apenas um dos tipos de critérios que pode ter sido empregado. Outros critérios também poderiam ter sido utilizados de maneira semelhante. Entretanto, o critério criado para o presente estudo foi o mais adequado até então.

4.2 Terrenos irrigados

Para o caso dos dados de terrenos irrigados, as duas primeiras componentes principais foram significativas. O gráfico de dispersão com base na avaliação sensorial e na classificação do modelo, para a primeira componente principal (Escore1) são ilustrados nas Figuras 4.21 e 4.22, respectivamente.

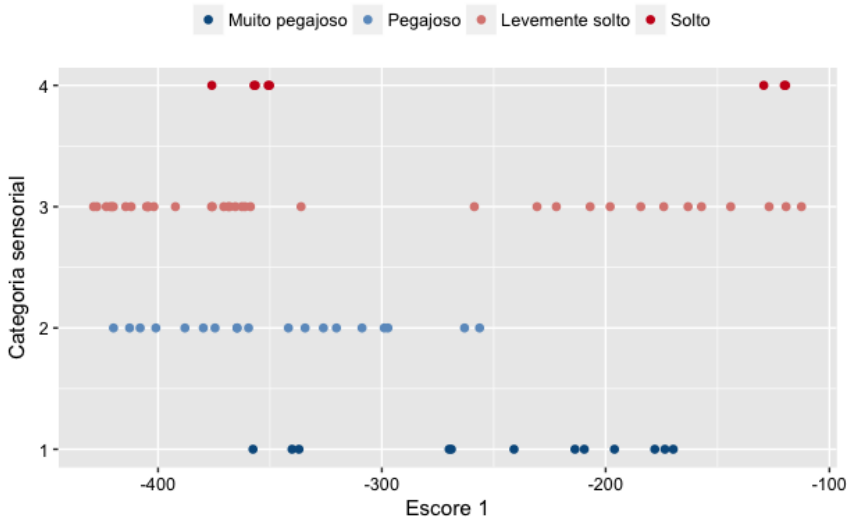

Figura 4.21: Gráfico de dispersão segundo classificação real dos dados, para terrenos irrigados.

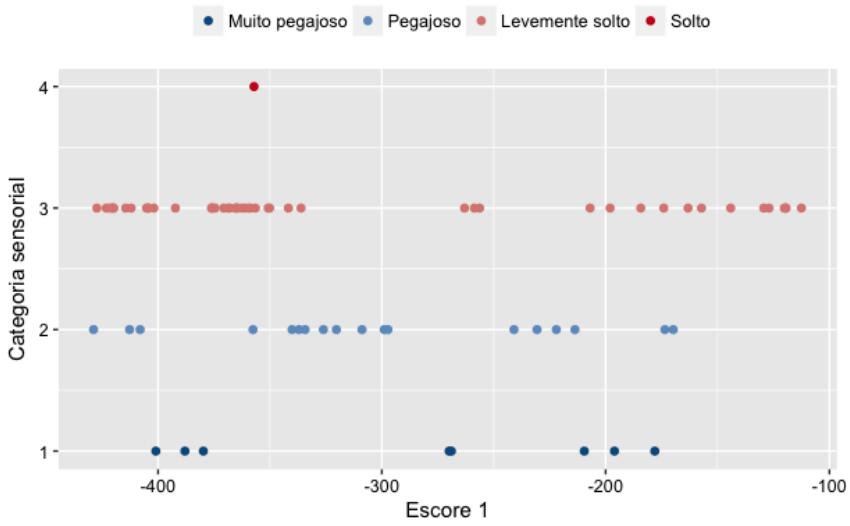

Figura 4.22: Gráfico de dispersão segundo classificação prevista dos dados, para terrenos irrigados.

Por meio dos gráficos de dispersão apresentados, nota-se que a classificação dos terrenos irrigados é ruim para ambas as situações. A distinção entre as transições entre as categorias não é clara nos gráficos apresentados. Após o estudo conduzido por Rios, observou-se um erro de classificação de 37% para os dados de terrenos irrigados [1], obtido por meio da Tabela 4.24.

Tabela 4.24: Classificação sensorial de pegajosidade do arroz versus a classificação prevista, para terrenos irrigados.

		Classe prevista			
	Categoria	MP	P	LS	S
Classe real	MP	5	7	0	0
	P	3	8	8	0
	LS	0	3	33	0
	S	0	0	7	1

Os limites onde a sobreposição dos dados entre categorias adjacentes, com base na dispersão real dos dados, são apresentos na Tabela 4.25.

Tabela 4.25: Limites das regiões de incertezada baseadas nos Escores1, para terrenos irrigados.

Região de incerteza	Limite inferior	Limite superior
S - LS (1)	-400,9847	-169,7856
LS - P (2)	-428,8197	-222,0386
P -MP (3)	-375,9832	-119,5907

Assim, são plotadas as barreiras no gráfico de dispersão, ilustradas na Figura 4.23.

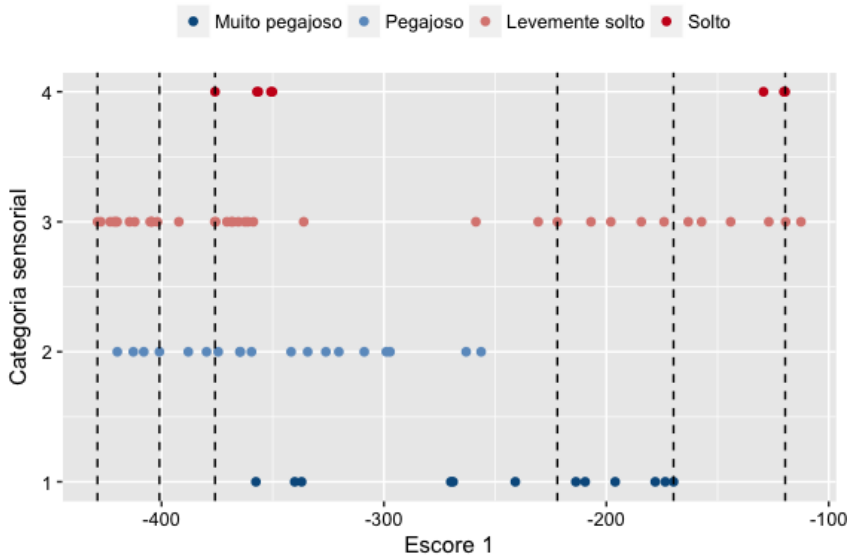

Figura 4.23: Limites das regiões de incerteza segundo classificação real dos dados, para terrenos irrigados.

Dessa maneira, consideradas as regiões de incerteza, o erro de classificação é de 0%. Entretanto, o custo resulta em 96% das observações não classificadas. Assim, com base na qualidade dos dados para terrenos irrigados, não é possível conduzir o estudo das regiões de incerteza.

4.3 Comparação entre tipos de barreiras

Após a análise de todos os critérios para definição das regiões deste estudo, é possível realizar a comparação entre estes, apresentada na Tabela 4.26.

Tabela 4.26: Comparação entre resultados segundo tipo de barreira empregado.

Barreira	Erro de classif.	Custo	Erro prop.	Classif. correta RI
Modelo	23,61%	0%	100%	66,67%
BER	17,07%	43,06%	58,82%	67,74%
BGI	17,91	6,94%	29,41%	0%
BEM	0%	70,83%	100%	66,67%
BPF	12,96%	25%	58,82%	44,44%

Assim, é possível ordenar os resultados obtidos, de acordo com o erro de classificação. Apresentados na Tabela 4.27.

Tabela 4.27: Comparação entre resultados segundo tipo de barreira empregado.

Barreira	Erro de classif.	Custo	Erro prop.	Classif. correta RI
BEM	0%	70,83%	100%	66,67%
BPF	12,96%	25%	58,82%	44,44%
BER	17,07%	43,06%	58,82%	67,74%
BGI	17,91	6,94%	17,65%	0%
Modelo	23,61%	0%	100%	76,39%

Por meio da Tabela 4.27, foi possível observar que, para erros proporcionais maiores, ou seja, para regiões de incertezas que continham maior número de erros de classificação (comparadas ao total), resultaram em custos mais elevados. Dessa maneira, mais amostras não são classificadas (sendo consideradas como incertezas).

É possível observar que o custo para obtenção de nenhum erro de classificação (0%) é muito alto, em que 70,83% das observações não são classificadas, uma vez que muitas observações que foram classificadas corretamente estavam dentro dos limites estabelecidos.

Assim, as medidas apresentadas servem como auxílio para a tomada de decisão do avaliador, sendo função deste escolher qual dos critério deseja utilizar, de acordo com o seu objetivo.

4.4 Resultados práticos para Embrapa - Shiny

O presente trabalho é resultado de uma cooperação com a Embrapa, visando à obtenção de resultados que possam ter utilidade prática para esta. Buscando a otimização e o fácil acesso aos resultados deste trabalho, foi desenvolvido um aplicativo interativo, junto ao grupo de extensão criado pelo professor orientador deste trabalho, voltado para o estudo de dados de arroz e feijão. O shiny é um ambiente para desenvolvimento de aplicações web usando a linguagem de programação R, que permite a criação de aplicativos interativos e de fácil manuseio pelo usuário. O estudante Rafael da Silva Lins (7° semestre — Bacharelado em Estatística) foi o principal programador que atuou no desenvolvimento do aplicativo.

Conforme explicitado detalhadamente neste trabalho, a principal motivação da criação do modelo de previsão das amostra de arroz é a redução de custos e otimização do processo para a análise de classificação das amostras de arroz. Dessa maneira, por meio de medidas instrumentais de viscosidade, obtidas via ferramentas laboratoriais, é possível fazer a substituição do processo de avaliação sensorial de arroz.

A interface desenvolvida tem como função fazer a automatização da classificação do arroz. Assim, as cinco medidas intrumentais de viscosidade, obtidas laboratorialmente, são inseridas no aplicativo, de modo que o aplicativo faz a previsão da classificação dessa nova amostra, com base no modelo criado por Rios e nas regiões de incerteza definidas ao longo deste estudo. O aplicativo fornece as probabilidades da amostra inserida pertencer a todas as categoria, os escores preditos por meio do modelo e se a amostra está na região de incerteza ou não, de acordo com o critério de definição das barreiras escolhido pelo avaliador. Além disso, um aplicativo interativo para registro de dados coletados foi desenvolvido. O aplicativo permite o usuário inserir o nome do avaliador que está fazendo a classificação da amostra, e faz o registro da classe em que a amostra foi alocada, diretamente em uma planilha Excel.

A Embrapa está fazendo uma nova coleta, em que um novo banco de dados com as 5 medidas intrumentais e a análise sensorial também é feita. Espera-se que a qualidade dos dados das novas amostras colhidas seja maior, de modo que o modelo criado possa fazer uma previsão ainda melhor da classificação do arroz, e que os resultados obtidos por meio desse trabalho possam ser otimizados.

Por fim, as Figuras 4.24 e 4.25 mostram os aplicativos desenvolvidos.

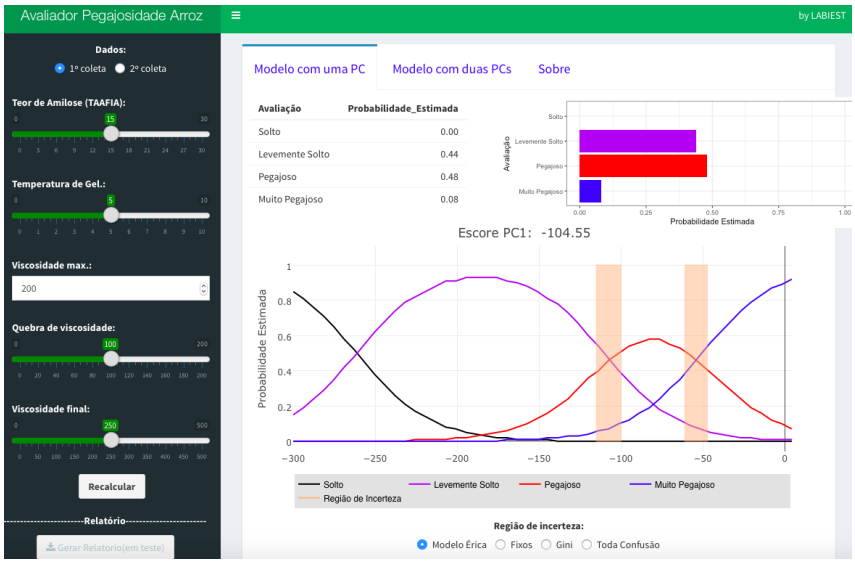

Figura 4.24: Aplicativo Shiny para Estudo de Regiões de Incerteza na Avaliação e Ajuste de Escalas de Classificação Sensorial de Arroz.

Figura 4.25: Aplicativo Shiny para registro de Classificação Sensorial de amostras de arroz.

5 Conclusão

Por meio dos resultados obtidos, foi possível observar que o método com base na dispersão real dos dados (BER), e o método que considera a classificação real versus a prevista (BEM) são os melhores critérios para a identificação dos erros de classificação. Em contrapartida, o primeiro método possui um custo muito elevado (43,06%), de modo que muitas amostras classificadas corretamente são consideradas como incerteza, de modo que o erro de classificação não reduziu muito (resultando em 17,07%) se comparado ao erro encontrado originalmente (23%). Em relação ao segundo critério, um erro de 0% foi estabelecido, entretanto aproximadamente 70% das amostra não são classificadas. Assim, os tipos de definição de incerteza que incluíram um número maior de erros de classificação, resultaram em custos com valores mais elevados, ou seja, mais amostras foram consideradas como incerteza.

Para essas regiões em que o custo é muito elevado, é mais vantajosa a retirada apenas da região que traz maior impacto no erro, de modo que o erro e o custo de classificação foram menores do que no caso em que ocorre a retirada das três regiões estabelecidas. Os resultados obtidos para esses dois métodos (BER e BEM) foram muito semelhantes, em que, por motivos de comparação, para o segundo caso mantiveram-se as regiões que retiravam todas as incertezas.

Para critérios que a identificação dos erros de classificação é menor, como no caso do Índice de Gini (BGI) e nas barreiras fixas em torno do ponto p_i de encontro entre as curvas de probabilidade (BPF), foi possível observar que o custo resultou em valores menores. Para o método BPF foi obtido o segundo menor erro de classificação (12,96%), uma vez que um número menor de amostras corretamente foram considerados como incerteza. Além disso, o método BGI resultou em um erro de classificação muito semelhante ao do BEM (no valor de 17,91%), com custo de classificação menor (25%).

Foi possível concluir que os critérios que levaram aos resultados mais vantajosos, em relação ao erro e custo de classificação, foram os critérios com base nas barreiras fixas e no índice de Gini. Assim, barreiras que identificam uma maior quantidade de erros de classificação possuem um custo muito elevado, fazendo com que o erro não reduza muito se comparado ao erro de classificação encontrado originalmente. Apesar de barreiras de incerteza que identifcam uma menor quantidade de erros, o custo associado à elas é menor, fazendo com que o erro de classificação reduza de melhor maneira se comparado ao erro

de classificação originalmente observado.

Por fim, o cenário ideal é que as regiões de incerteza estejam em torno dos pontos p_i de transição entre as curvas de probabilidade de categorias adjacentes. Após o estudo conduzido, pode-se observar que o critério com base na dispersão real, e com base nas classificações errôneas entre o sensorial e o modelo, resultou em regiões de incerteza distantes dos pontos de transição entre as curvas de probabilidade, ou seja, eles possuem seus limites com amplitudes maiores. Isso pode ser um indício de que o modelo ainda não está prevendo da melhor maneira a classificação das amostras, ou de que a avaliação sensorial ainda deve ser melhorada.

Com base no critério BPF, pode-se concluir que aproximadamente 58% dos erros de classificação se encontram próximos das regiões de transição entre as categorias vizinha.

Considera-se que ainda existem melhorias a serem conduzidas por parte da avaliação sensorial do arroz conduzida na Embrapa e na efetividade da previsão do modelo para que assim, o estudo da previsão da classificação do arroz possam ser extendidos para os dados da variável de dureza, e para os dados de terrenos irrigados.

Por fim, é função do avaliador definir qual método para definição de incerteza ele deseja utilizar, com base nos resultados apresentados.

6 Bibliografia

Referências

[1] Rios, E. S. (2015). *Modelos Estatísticos para Avaliação da Qualidade Culinária de Arroz: Textura e Propriedades Viscoamilográficas.* Trabalho de conclusão de curso, Departamento de Estatistica, UnB.

[2] de Oliveira, G. S. (2015). *Modelos de Regressão com Resposta Ordinal para Avaliação de Textura de Arroz.* Trabalho de conclusão de curso, Departamento de Estatística, UnB.

[3] Okabe, M. (1979). *Texture measurement of cooked rice and its relationship to the eating quality.* J, texture study.

[4] Bueno, P. D. F. (2008). *Viscoamilografia na estimativa do teor de amilose e caracteristicas de consumo de arroz.* Universidade Federal de Pelotas.

[5] Agresti, A. (2002). *Categorical Data Analysis.* Wiley, segunda edição.

[6] Kleinbaum, D. G. ; Klein, M. (2010). *Logistic regression: A Self-learning Text.* 3. ed. Spinger. New York.

[7] Ferreira, C. M.; Pinheiro, B. S.; de Souza, I. S. F. e de Morais, O. P. (2005). *Qualidade do Arroz no Brasil: Evolução e Padronização.* Embrapa.

[8] Kutner, M. H.; Nachtsheim, C.; Neter, J. (2005). *Applied Linear Statistical Models.* Boston: McGraw-Hill Irwin.

[9] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statisitical Analysis.* sixth ed. Upper Saddle River, NJ: Pearson/Prentice Hall.

[10] Weisberg, S (2005). *Applied Linear Regression.* third ed. Wiley, NJ: Hoboken.

[11] Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression.* second ed. John Wiley 7 Sons, 2000.

[12] Hartshorn, S. (2016). *Machine Learning with Random Forests and Decision Trees.* Amazon Digital Services.

[13] Yeung, K. Y.; Ruzzo, W. L. (2001). *Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data.* Bioinformatics, Volume 7, pages 763-774.)

[14] Hubert, L. and Arabie, P. (1985) *Journal of Classification. Volume 2, Number 1, Page 193.*

[15] Xie, Y. (2015). *Dynamic Documents with R and Knitr. CRC Press, segunda edicao.*