Universidade Federal do Espírito Santo
Departamento de Estatística
Análise de Dados Categorizados - Trabalho Final

Um Modelo de Classificação para comestibilidade de Cogumelos

Aluno: Franco Lovatti Souza Pinto
Professora: Nataly Adriana Jimenez Monroy

Junho
2024

Conteúdo

1 Resumo

Este relatório tem como propósito desenvolver um modelo de classificação para a variável de interesse *Comestibilidade*, possuindo uma introdução do tema e os dados que foram estudados, uma análise exploratória, construção dos modelos de classificação e apresentação do desempenho destes modelos. Este documento possui caráter de trabalho final da disciplina de Análise de Dados Categóricos.

Os dados abordados no estudo são da **base de dados "Mushroom"**`https://archive.ics.uci.edu/dataset/73/mushroom`, advinda das informações do livro "Field Guide to Mushrooms", em português "Guia de Campo para Cogumelos", da National Audubon Society. A National Audubon Society (https://www.audubon.org/) é uma organização não governamental americana de conservação da natureza que possui uma série de Guias(livros) sobre espécies da natureza. Os dados que são objeto de estudo deste trabalho são provenientes de um destes Guias produzidos pela Audubon.

2 Introdução

Pertencentes ao Reino Fungi, os cogumelos são, na verdade, o "nome popular"de uma parte do corpo de diversos fungos pertencentes aos filos *Ascomycota* e *Basidiomycota*. É o corpo frutífero, formado por várias hifas, responsável por produzir esporos que se espalham pelo ambiente e mantêm o ciclo reprodutivo destas milhares de espécies de fungos.

Os cogumelos possuem diferentes características utilizadas no processo de identificação. Este relatório faz análises preliminares de 23 destas características de cogumelos da família *Agaricus and Lepiota*, sintetizadas em 23 variáveis categóricas que descrevem 8124 observações da base de dados **"Mushroom"**, advinda do livro "Field Guide to Mushrooms"produzido pela organização americana National Audubon Society (https://www.audubon.org/), apresenta modelos logísticos ajustados com objetivo de prever se um cogumelo é comestível ou não, e mostra os resultados de previsão testados com uma parte dos dados.

3 Metodologia

Inicialmente, foram feitas uma introdução e uma análise exploratória dos dados, calculando medidas de associação entre as variáveis explicativas e a variável de interesse *Comestibilidade*, e apresentando gráficos que expõem visualmente possíveis associações.

Como os dados se encaixam em um tipo de estudo transversal, uma estatística de teste utilizável é a estatística Q_p Qui-Quadrado, portanto qualquer medida de associação que se baseia na estatística Q_p é interessante para descrever a relação entre as variáveis dos dados. Desta maneira, o primeiro passo dado foi calcular a associação entre a variável de interesse *Comestibilidade* e as outras 22 variáveis explicativas, por meio do Coeficiente de Contingência de Pearson Ajustado, representado por C^*, que consegue traduzir a possível associação com um valor entre 0 e 1, sendo 0 nenhuma associação e 1 máxima associação.

Em seguida, com base nestas medidas, as 5 variáveis explicativas com maior relação com a variável *Comestibilidade* foram selecionadas e gráficos que apresentam visualmente esta associação foram construídos para melhor entendimento das tendências dos dados. Adiante, levando em mente o objetivo de criação de um modelo de classificação, é interessante a verificação de possíveis correlações entre covariáveis, e a mesma medida de associação exposta acima foi utilizada para verificar isso.

Posteriormente, no momento de ajuste dos modelos, a técnica de validação cruzada foi utilizada, atrelada à medida de desempenho dos modelos, que foi a área abaixo da Curva Característica de Operação do Receptor (Curva ROC). Por fim, o tratamento dos dados realizado, juntamente com a utilização de penalizações para os coeficientes da Regressão Logística, mitigaram os impactos da multicolinearidade e geraram resultados satisfatórios nas previsões da parte de Teste.

4 Análise Exploratória dos Dados

4.1 Conjunto de Dados

A base de dados **"Mushroom"**, objeto de estudo deste relatório, possui 23 variáveis categóricas, com diferentes números de categorias para cada uma delas, e 8124 observações. A variável de interesse deste estudo será a variável *Comestibilidade*, que indica se o cogumelo observado, 1 dos 8124 possíveis, é comestível ou venenoso. Portanto, a variável de interesse *Comestibilidade* é dicotômica: {Sim ou Não}. Na Tabela 1 há informações sobre as variáveis.

Variáveis da Base de Dados "Mushroom"		
Nome da Variável	Tipo	Número de Categorias
Comestibilidade	Dicotômica: sim ou não	2
Formato do Chapéu	Nominal	6
Superfície do Chapéu	Nominal	4
Cor do Chapéu	Nominal	10
Machucados	Dicotômica: sim ou não	2
Odor	Nominal	9
Acessório das Guelras	Nominal Dicotômica	2
Espaçamento das Guelras	Nominal Dicotômica	2
Tamanho das Guelras	Nominal Dicotômica	2
Cor das Guelras	Nominal	12
Formato do Caule	Nominal Dicotômica	2
Raiz do Caule	Nominal	5
Superfície do Caule Acima do Anel	Nominal	4
Superfície do Caule Abaixo do Anel	Nominal	4
Cor do Caule Acima do Anel	Nominal	9
Cor do Caule Abaixo do Anel	Nominal	9
Tipo de Véu	Nominal	1
Cor do Véu	Nominal	4
Número de Anéis	Ordinal (0, 1, 2)	3
Tipo do Anel	Nominal	5
Cor da Impressão de Esporos	Nominal	9
População	Nominal	6
Habitat	Nominal	7

Tabela 1: Variáveis da Base de Dados Mushroom

É possível verificar, pelo gráfico abaixo, que a proporção de cogumelos **comestíveis** e **não comestíveis** é bem equilibrada. O número de cogumelos comestíveis existentes na base de dados é 4208, o que representa $51,8\%$, e o número de cogumelos não comestíveis é de 3916, representando os outros $48,8\%$.

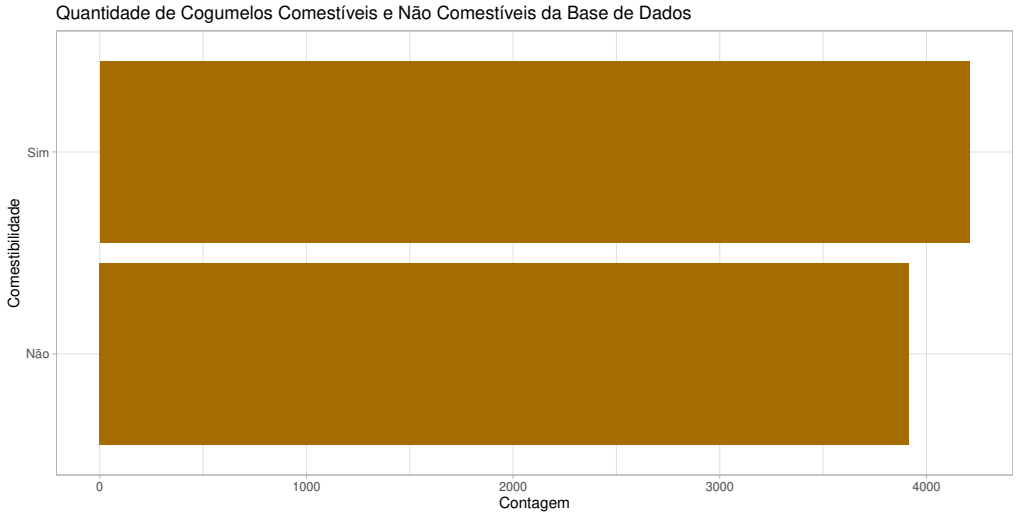

Gráfico 1: Quantidade de cogumelos em cada categoria da variável Comestibilidade

4.2 Associações com a Variável de Interesse

Em primeiro momento, para se ter uma noção inicial das associações mais fortes das variáveis explicativas com a variável de interesse *Comestibilidade*, foi calculado o Coeficiente de Contingência de Pearson em sua forma ajustada, dado por $C^* = \sqrt{\frac{Q_p}{Q_p+n}} \cdot \sqrt{\frac{k}{k-1}}$, sendo Q_p a estatística Qui-Quadrado calculada a partir da tabela de contingência montada com duas variáveis e k o valor mínimo entre o número de linhas e o número de colunas desta tabela de contingência. Foram selecionadas as 5 associações mais fortes com a variável *Comestibilidade*, segundo o coeficiente C^*, e expostas na tabela abaixo.

Associação das Variáveis Explicativas com a variável *Comestibilidade*	
Variáveis Explicativas	Coeficiente de Contingência de Pearson Ajustado C^*
Odor	0.9851829
Cor da Impressão de Esporos	0.8504393
Cor das Guelras	0.7958898
Tipo do Anel	0.7305176
Superfície do Caule Acima do Anel	0.7167706

Tabela 2: Associação das Variáveis Explicativas com a variável *Comestibilidade*

4.3 Gráficos - Visualizar Associações entre Variáveis

A partir das variáveis com maior associação com a variável de interesse foram construídos alguns gráficos para explicitar visualmente estas possíveis relações.

O primeiro destes gráficos que apresentam as mais fortes associações é o Gráfico 2, apresentado abaixo, relacionando as variáveis *Odor* e *Cor da Impressão dos Esporos* com a variável de interesse *Comestibilidade*. A "Esporada"ou Impressão de Esporos é um importante processo experimental para a identificação de cogumelos. Este processo consiste em espalmar a superfície produtora de esporos do cogumelo em uma folha branca ou preta e verificar a cor da impressão gerada sobre a folha. A variável *Cor da Impressão dos Esporos* se refere a este experimento descrito. Já a variável *Odor* se refere ao odor emitido pelo cogumelo.

O Gráfico 2 deixa bem evidente a capacidade da variável *Odor* em discriminar as categorias da variável *Comestibilidade*. A única categoria de *Odor* que possui cogumelos tanto comestíveis, quanto não comestíveis é a categoria "Sem cheiro". É perceptível também a relação de *cores de impressão de esporos* para cada categoria de *Odor* e de *Comestibilidade*. Existem cores bem predominantes em certas categorias.

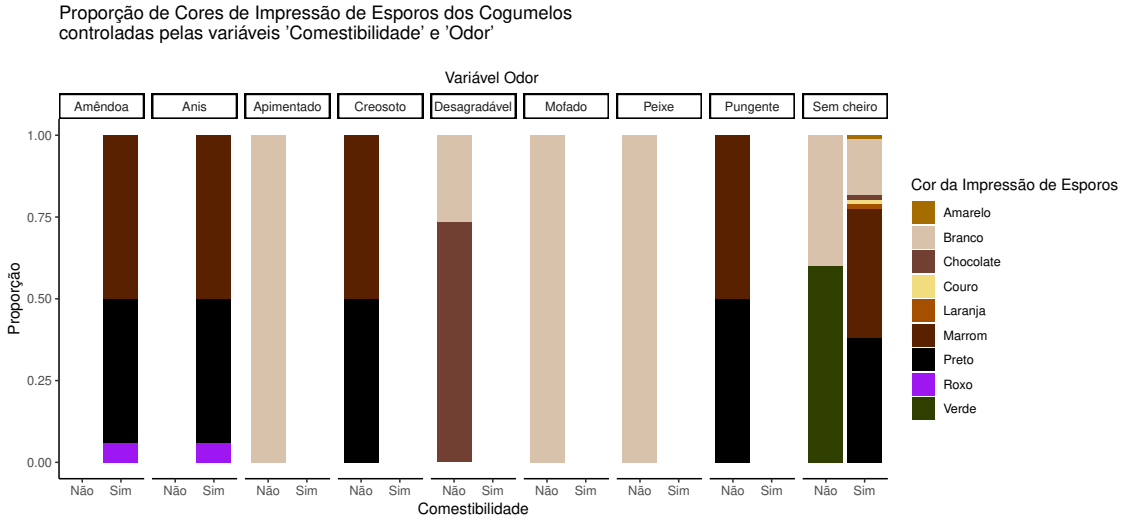

Gráfico 2: Gráfico de barras que mostra a proporção de Cores de Impressão de Esporos, controladas por Comestibilidade e Odor

Adiante, o Gráfico 3 possui a mesma estrutura e ideia do Gráfico 2, desta vez com a proporção de cores das guelras dos cogumelos, representada pela variável *Cor das Guelras*, controladas pelas variáveis *Comestibilidade* e *Tipo de Anel*. Neste caso, é possível perceber uma aparente fraca associação para algumas categorias de cores e já para outras vê-se uma possível forte associação. Para os tipos de anel, 3 das 5 categorias discriminam completamente a variável de interesse, já outras duas categorias, possuem cogumelos tanto comestíveis, quanto não comestíveis.

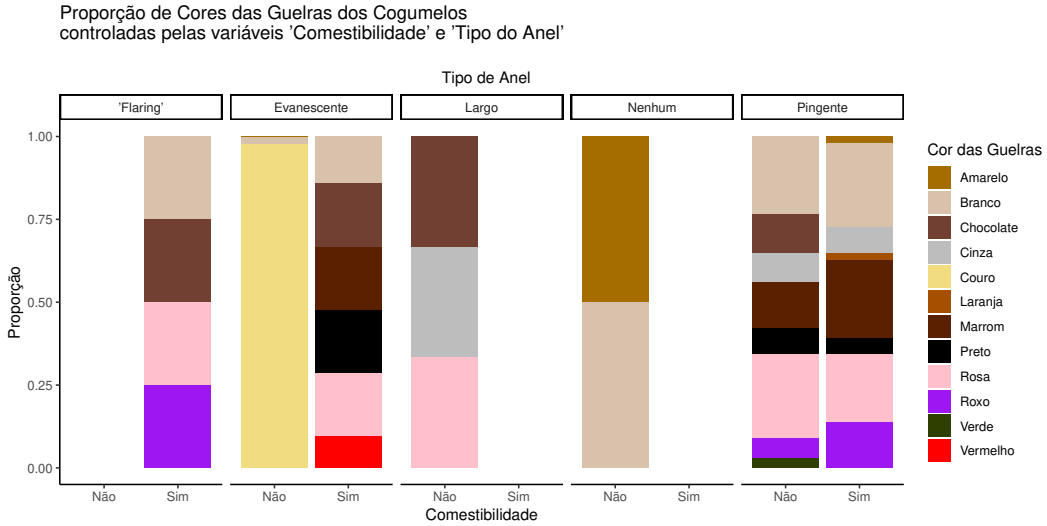

Gráfico 3: Gráfico de barras que mostra a proporção de Cores das Guelras, controladas por Comestibilidade e Tipo de Anel

Por fim, o Gráfico 4 evidencia a diferença de proporções de cogumelos comestíveis e não comestíveis para a categoria "Sedosa"da variável *Superfície de Caule acima do Anel do Cogumelo*, insinuando uma forte relação com a variável de interesse.

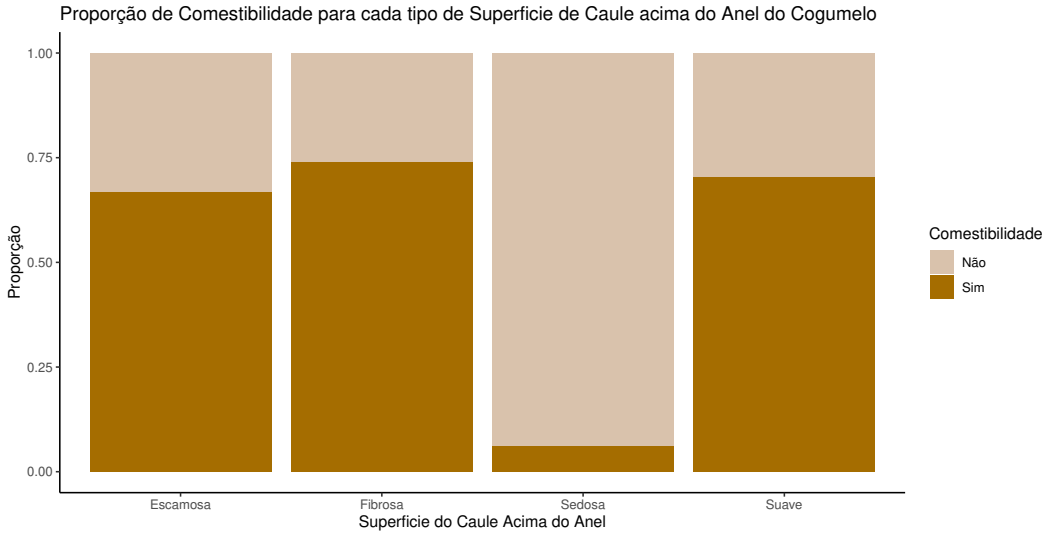

Gráfico 4: Gráfico de barras que mostra a proporção de cogumelos comestíveis para cada tipo de Superfície do Caule acima do Anel

As análises e gráficos feitos fortalece a crença de que as 5 variáveis com o Coeficiente de Contingência de Pearson Ajustado mais altos, trabalhadas no relatório, são importantes para o processo de modelagem para classificação de cogumelos como comestível ou não comestível.

4.4 Associações entre Covariáveis

Tendo em vista o objetivo de construir um modelo de classificação, é interessante a verificação de possíveis associações entre as covariáveis. Foi construído um gráfico em formato de "mapa de calor"para expor estas relações entre as 5 covariáveis com maior associação com a variável de interesse. O coeficiente utilizado para esta etapa foi também o Coeficiente de Contingência de Pearson Ajustado.

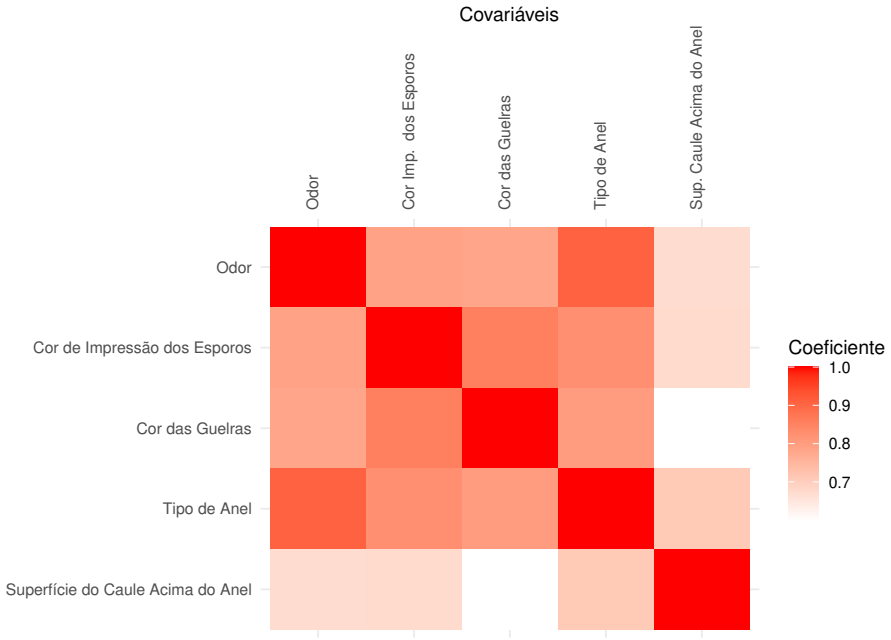

Gráfico 5: Mapa de Calor das associações entre 5 das covariáveis do conjunto de dados

É evidente a existência de multicolinearidade. Muitas das possíveis covariáveis para ajustar o modelo são correlacionadas.

5 Construção do Modelo de Classificação

Finalizada a análise exploratória das variáveis contidas na base de dados "Mushroom", é de interesse a criação de um modelo de classificação para a variável *Comestibilidade*.

5.1 Divisão da base de dados

Antes da criação dos modelos, a base de dados foi dividida em duas partes distintas: uma parte para ajustar os modelos que foi chamada de Parte de Ajuste (60% dos dados) e uma outra parte para testar os modelos e verificar seus poderes preditivos que foi chamada de Parte de Teste (40% dos dados).

Com o intuito de ajustar um modelo com alto poder preditivo e evitar sobre-ajuste, foi utilizada a técnica de validação cruzada denominada "k-fold". Neste caso, foi utilizado $k = 5$. Esta técnica consistiu em subdividir a Parte de Ajuste em $k = 5$ partes mutuamente exclusivas e de mesmo tamanho em que $k - 1 = 4$ destas partes são utilizadas para estimar os parâmetros e a

outra parte utilizada para testar o modelo. Este processo é realizado $k = 5$ vezes de forma iterativa, de maneira que todas as partes servem como teste em uma das iterações, e ao final de cada iteração é calculada uma medida de desempenho do modelo.

O objetivo desta técnica é evitar a possibilidade de ajustar modelos com pouco poder de prever dados novos.

5.2 Área Abaixo da Curva ROC

A medida de desempenho calculada para cada modelo ao final de cada uma das $k = 5$ iterações foi a área abaixo da Curva Característica de Operação do Receptor (Curva ROC), **AUC**, *"Area Under Curve"*. A área média entre as 5 áreas de um modelo para cada uma das 5 iterações é a utilizada como comparação entre os modelos. Os modelos com maiores área média foram selecionados.

5.3 Tratamento dos Dados

Alguns passos de tratamento dos dados foram efetuados antes de ajustar os modelos:

1. Variáveis com variância 0 foram removidas
2. Variáveis com variância próximas de 0 também foram removidas.

   Observação: a frequência de corte para determinar variância próxima de 0 foi de 5% ou seja a razão entre a o número de observações da 2ª categoria mais comum e o total de observações ser menor ou igual a 0.05.
3. Variáveis categóricas foram transformadas em variáveis indicadoras

5.4 Modelos ajustados

Tendo em vista o objetivo de classificar a variável resposta *Comestibilidade*, que possui resposta binária, e também considerando o fato de que há multicolinearidade nos dados, foram ajustados e testados modelos de Regressão Logística, como também modelos de Regressão Logística com Penalização. Dois tipos de penalização foram considerados: $L1 = \lambda \sum_{j=1}^{p} |\beta_j|$ e $L2 = \lambda \sum_{j=1}^{p} \beta_j^2$. O modelo de Regressão Logística com Penalização possui função log-verossimilhança análoga a um modelo de Regressão Logística com p parâmetros, somada uma parcela $L1$ ou $L2$. Desta maneira, λ é um

novo parâmetro a ser estimado. Estas penalizações nos coeficientes β_j da Regressão podem reduzir o impacto da multicolinearidade.

5.4.1 Regressão Logística

Sendo $\mathbf{Y}$ um vetor de respostas para a variável de interesse e $\mathbf{X}_i$ um conjunto de variáveis explicativas para a resposta Y_i, $\pi(\mathbf{X}_i)$ é a probabilidade desconhecida de uma observação Y_i satisfazer uma característica de interesse dado os valores de um conjunto de variáveis explicativas representado por $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{i1-p})$. Desta forma, a distribuição da variável aleatória Y_i pode ser vista como um Distribuição Bernoulli:

Y_i	Probabilidade
1	$\Pr(Y_i = 1\|\mathbf{X}_i) = \pi(\mathbf{X}_i)$
0	$\Pr(Y_i = 0\|\mathbf{X}_i) = 1 - \pi(\mathbf{X}_i)$

Tabela 3: Distribuição de probabilidade da variável de interesse Y

O modelo de Regressão Logística múltipla considera, por meio da transformação *logit* que:

$$\pi(\mathbf{X}_i) = \Pr(Y_i = 1|\mathbf{X}_i)$$

$$= \frac{\exp\{\beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{ip-1}\}}{1 + \exp\{\beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{ip-1}\}} \tag{1}$$

Desta forma, considerando $\pi(\mathbf{X}_i) = \pi_i$, a função verossimilhança $L(\beta)$ é dada por

$$L(\beta_0, \beta_1, ..., \beta_{p-1}|y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

E, substituindo π_i por (1) a função log-verossimilhança $l(\beta)$ dada por

$$ln(L(\beta)) = l(\beta) = ln\{L(\beta_0, \beta_1, ..., \beta_{p-1}|y_1, y_2, ..., y_n)\}$$

$$= \sum_{i=1}^{n} y_i(\beta_0, \beta_1 x_{i1}, ..., \beta_{p-1} x_{ip-1}) - \sum_{i=1}^{n} ln\{1 + \exp(\beta_0, \beta_1 x_{i1}, ..., \beta_{p-1} x_{ip-1})\}$$

5.4.2 Regressão Logística com Penalização

Como explicado anteriormente, as funções log-verossimilhança dos modelos com penalização são análogos à função log-verossimilhança $l(\beta)$ descrita acima, porém com parcelas de penalização relacionadas aos p coeficientes e um novo parâmetro λ a ser estimado.
Com a parcela $L1$, a log-verossimilhança $l^{L1}(\beta)$ fica:

$$l^{L1}(\beta) = l(\beta) - L1$$

$$l^{L1}(\beta) = l(\beta) - \lambda \sum_{j=1}^{p} |\beta_j|$$

Com a parcela $L2$, log-verossimilhança $l^{L2}(\beta)$ fica:

$$l^{L2}(\beta) = l(\beta) - L2$$

$$l^{L2}(\beta) = l(\beta) - \lambda \sum_{j=1}^{p} \beta_j^2$$

5.5 Otimização do Ponto de Corte

Inicialmente, o ponto de corte foi selecionado de forma a maximizar a soma da sensibilidade + especificidade. O resultado dos modelos de classificação com este ponto de corte foi relativamente satisfatório mas houveram previsões de falsos positivos.

Como é importante que o modelo não cometa esse tipo de erro, pois não é interessante prever que um cogumelo é comestível, sendo ele venenoso ou não comestível, a maneira de seleção do ponto de corte foi alterada.

O ponto de corte final foi selecionado, maximizando a especificidade, dado que a sensibilidade teria que ser igual a 1.

5.6 Modelos Finais

Nesse sentido, ao utilizar a metodologia explicada de validação cruzada para evitar sobreajuste, o tratamento especificado, a medida de comparação de desempenho como sendo a área abaixo da curva ROC e os modelos de regressão logística com penalização apresentados, alguns modelos foram ajustados. O modelo de Regressão Logística sem penalização obteve problemas computacionais em sua estimação. Devido a isso, e sabendo da existência de multicolinearidade, modelos com penalização passaram a ser considerados.

O modelo com melhor desempenho, seguindo métrica de maior área abaixo da curva ROC, de Regressão Logística com a penalização $L2$ e com o ponto de corte sendo escolhido a partir da maximização de sensibilidade+especificidade, obteve os seguintes resultados com os 40% dos dados separados para teste (3251):

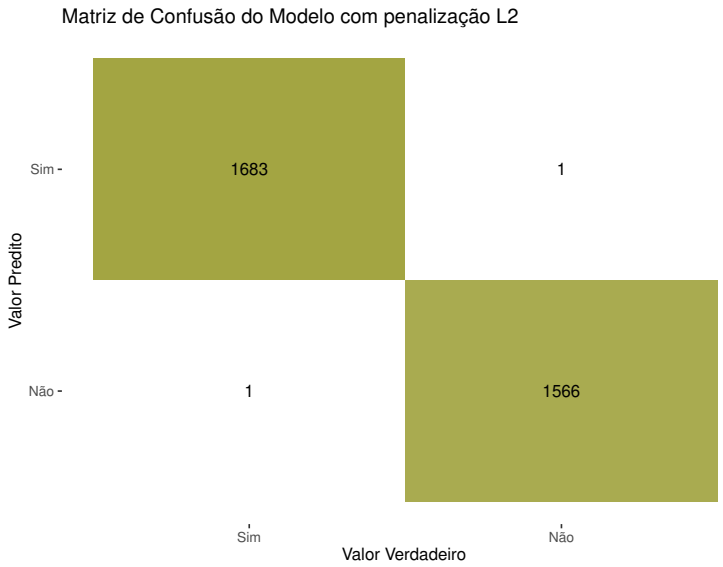

Gráfico 6: Matriz de Confusão do Modelo com penalização L2 e ponto de corte 1

Já o mesmo modelo com penalização $L2$, porém com o **ponto de corte sendo otimizado**, com a condição de que a sensibilidade tem que ser igual 1, obteve o seguinte resultado, sobre os mesmos 40% dos dados de teste:

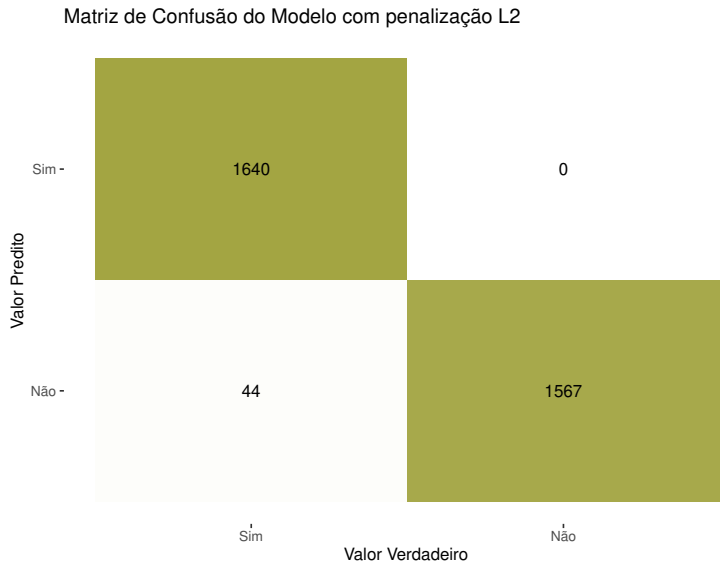

Gráfico 7: Matriz de Confusão do Modelo com penalização L2 e ponto de corte 2

O número de falsos positivos baixou para zero com este novo ponto de corte.

Já para o melhor modelo, seguindo a métrica de desempenho de maior área abaixo da curva ROC, porém desta vez com a penalização $L1$, a maneira de seleção do ponto de corte não teve impacto. Mesmo com o ponto de corte não sendo otimizado dado uma sensibilidade mínima, o modelo obteve assertividade de 100%. A matriz de confusão do Modelo com Penalização $L1$, para qualquer uma das duas formas de seleção de ponto de corte foi:

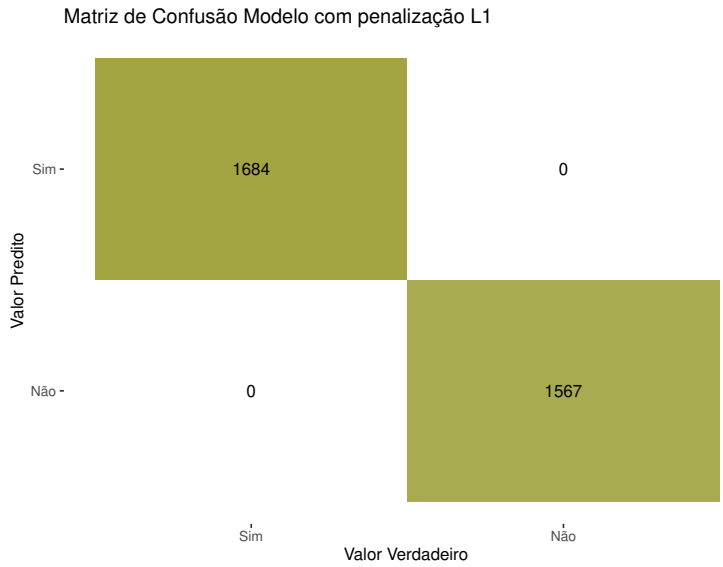

Gráfico 7: Matriz de Confusão do Modelo com penalização L1

6 Discussão e Conclusão

Tornou-se evidente que o modelo com a penalização do tipo $L1$ dada por $\lambda \sum_{j=1}^{p} |\beta_j|$ é mais eficiente para a classificação da variável *Comestibilidade* do conjunto de dados "Mushroom"estudado no relatório. Portanto o modelo final obteve as medidas sobre a parte de Teste dos dados dada pela tabela 4:

Medida	Valor
Prevalência	0.5179
VPP	1
VPN	1

Tabela 4: Medidas do Modelo Final sobre a parte de Teste

Referências

Evenson, Vera Stucky (1997). Mushrooms of Colorado and the Southern Rocky Mountains. [S.l.]: Big Earth Publishing.

Silva, L., & Fortuna, J. (2020). MACROFUNGOS ENCONTRADOS NO CAMPUS X DA UNIVERSIDADE DO ESTADO DA BAHIA.

Azul, Anabela Marisa. Cogumelos do Paul da Madriz. Imprensa da Universidade de Coimbra/Coimbra University Press, 2010.

Hastie, T., Tibshirani, R., Friedman, J. (2009). Linear Methods for Classification. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-84858-7_4

Doerken S, Avalos M, Lagarde E, Schumacher M. Penalized logistic regression with low prevalence exposures beyond high dimensional settings. PLoS One. 2019 May 20;14(5):e0217057. doi: 10.1371/journal.pone.0217057. PMID: 31107924; PMCID: PMC6527211.

NAKAMURA, Karina Gernhardt. Multicolinearidade em modelos de regressão logística. 2013. Dissertação (Mestrado em Estatística) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2013. doi:10.11606/D.45.2013.tde-28052013-222241. Acesso em: 2024-06-29.