Relatório — Classificação de Voz com Regressão Logística

Gerado automaticamente em 30/10/2025 20:59

# 1) Resumo executivo

* Base voice.csv por URL: 3168 amostras × 21 colunas (20 numéricas + rótulo).
* Classes balanceadas (0=male, 1=female): 50% / 50%.
* Split estratificado: treino (2534×20) e teste (634×20); padronização aplicada.
* GridSearchCV → melhor modelo L2, C=10.0, solver=liblinear, max\_iter=1000 (acurácia média CV=0.9767).
* Teste (logístico padrão): Acc=0.9621, Prec=0.9651, Rec=0.9590, F1=0.9620; calibração superior ao RMSE (Brier 0.0287 vs 0.0407).

# 2) Por que essa abordagem funcionou

• Padronização: as escalas variam muito (ex.: kurt std≈134.93 vs Q75 std≈0.0236). Sem padronizar, o solver pesaria variáveis maiores.

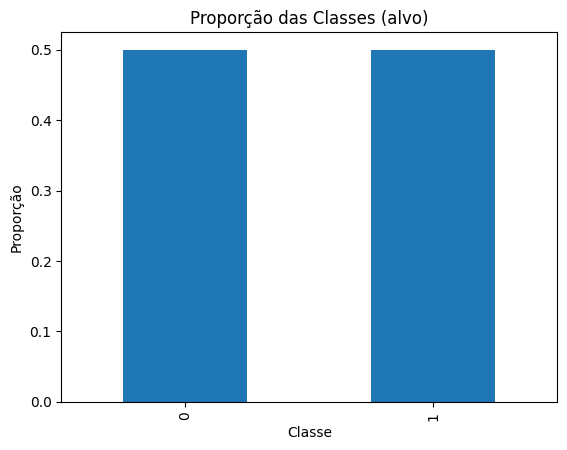
• Regularização: há multicolinearidade relevante (meanfreq×centroid r=1.000; skew×kurt r=0.977; median×meanfreq r=0.925; Q25×meanfreq r=0.911). L2 estabiliza coeficientes; L1 poderia selecionar, mas com C=100 não zerou.

• Validação cruzada: escolheu hiperparâmetros com boa generalização (desvios baixos; train≈test).

# 3) Parte A.1 — Exploração dos dados (EDA)

## 3.1 Balanceamento das classes

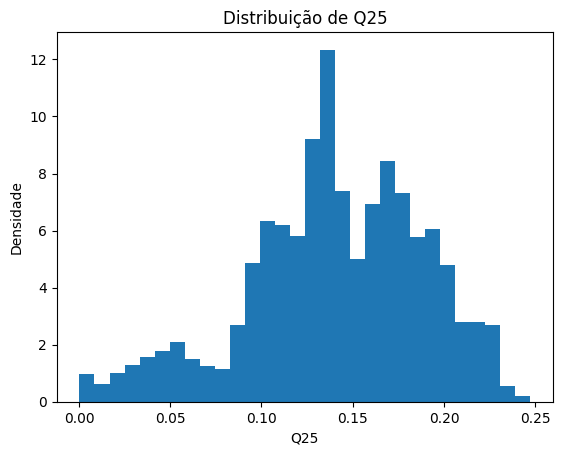
Conjunto perfeitamente balanceado (50/50). Isso evita viés trivial de acurácia e permite comparar Precisão/Recall/F1 sem distorção.



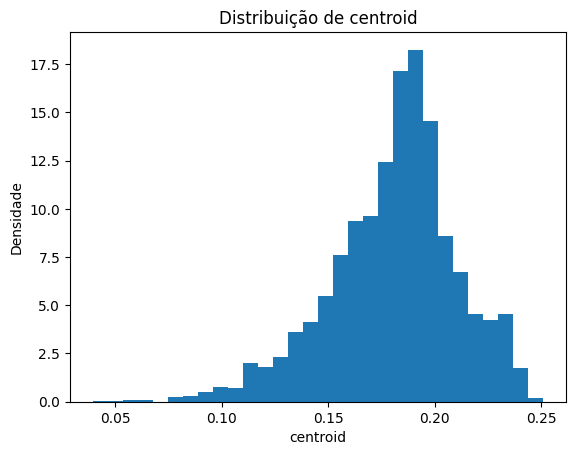
*Figura A1 — Proporção das classes (0=male, 1=female).*

## 3.2 Distribuições por variável

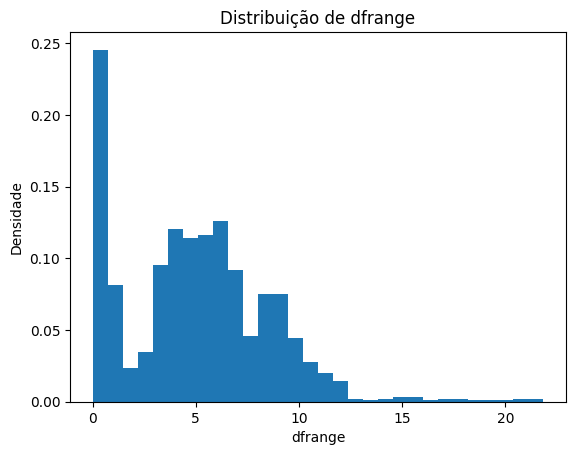
As distribuições por classe mostram separações claras em variáveis ligadas a frequência/energia. Esse padrão antecipa maiores |coef| nessas features no modelo logístico.



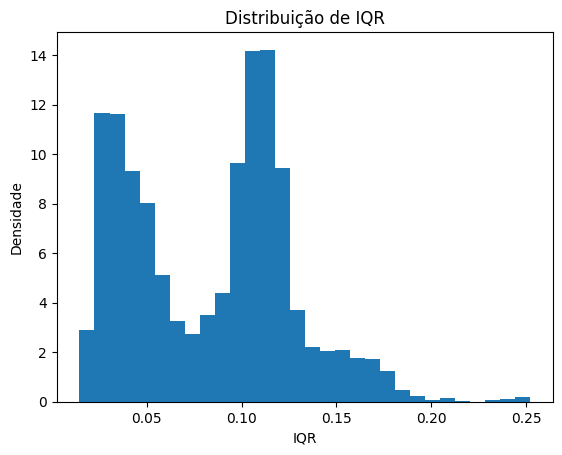
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 1)*



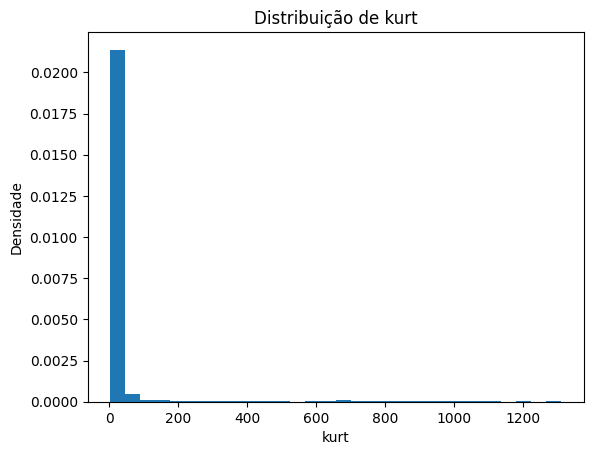
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 2)*



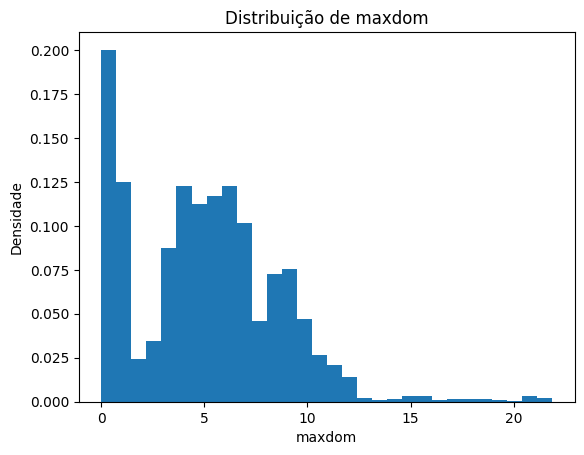
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 3)*



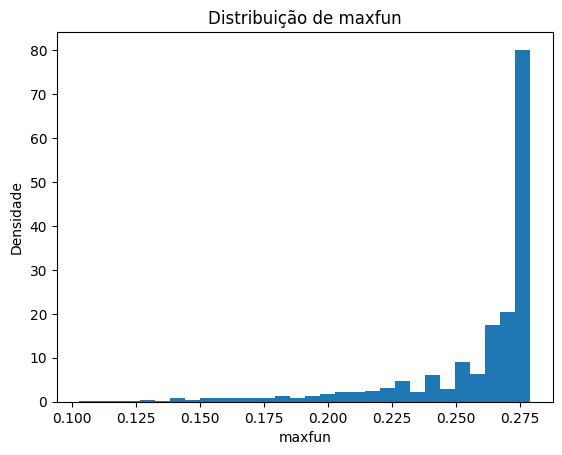
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 4)*



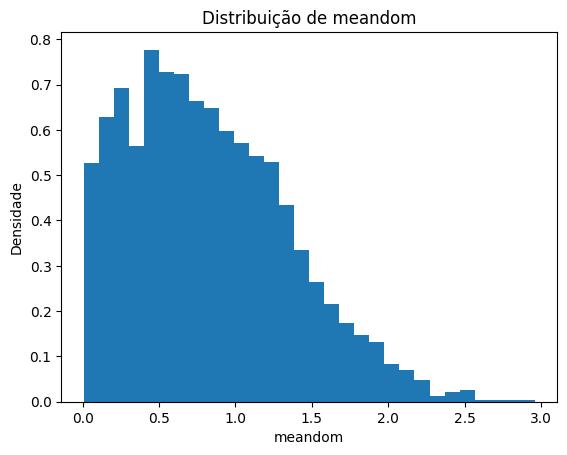
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 5)*



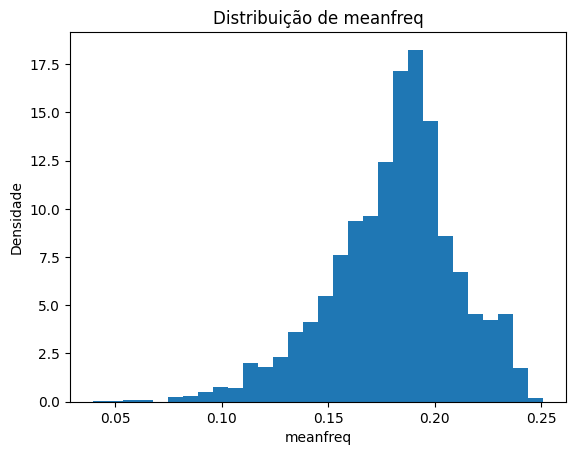
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 6)*



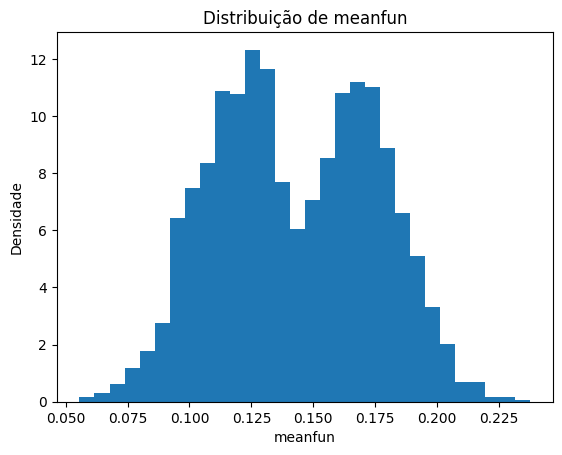
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 7)*



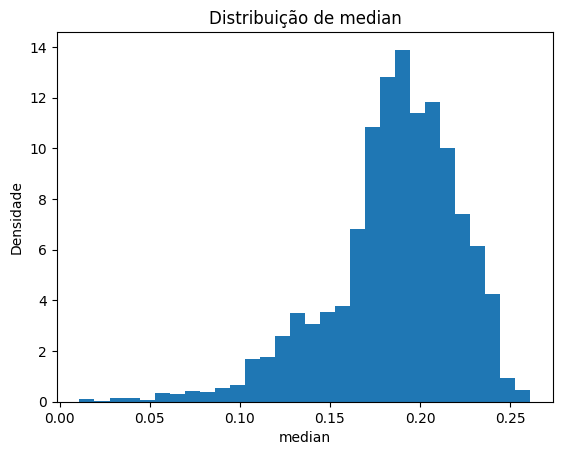
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 8)*



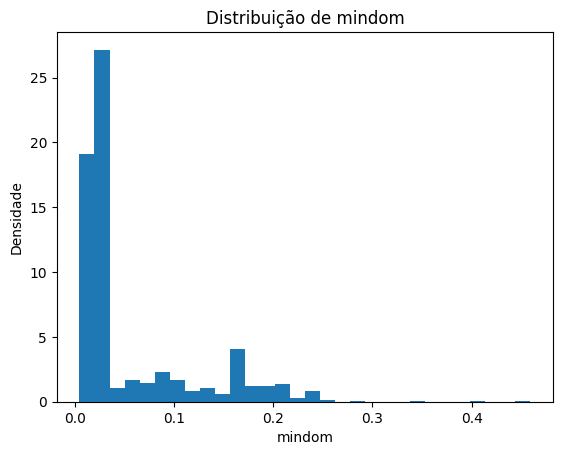
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 9)*



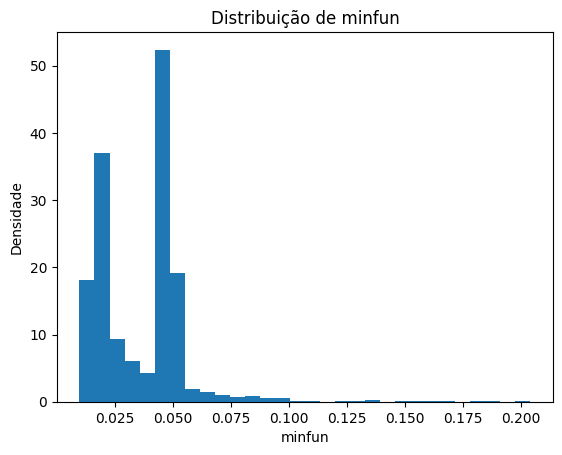
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 10)*



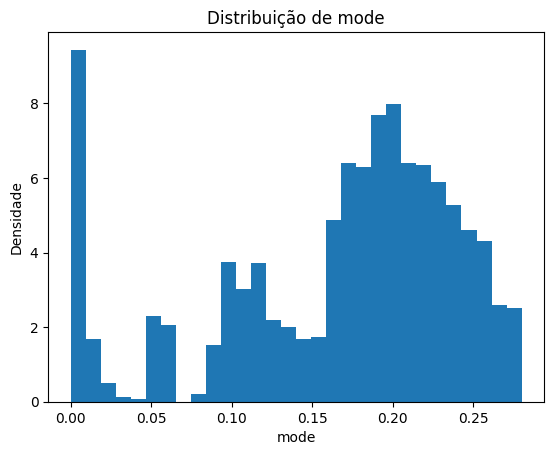
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 11)*



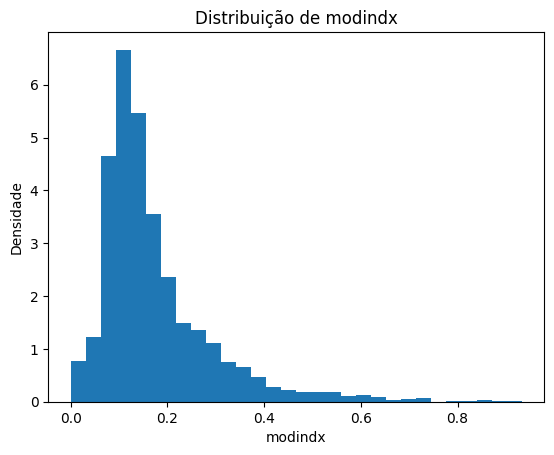
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 12)*



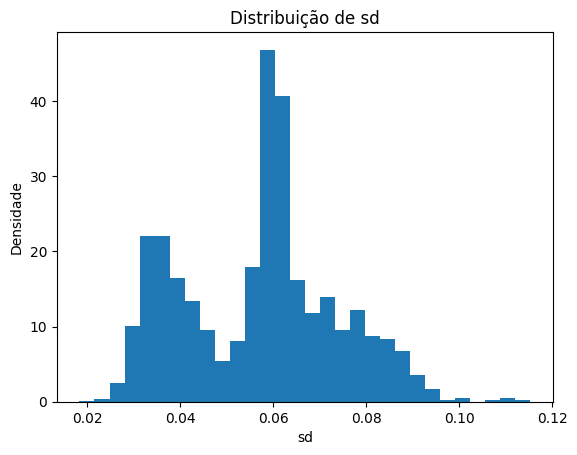
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 13)*



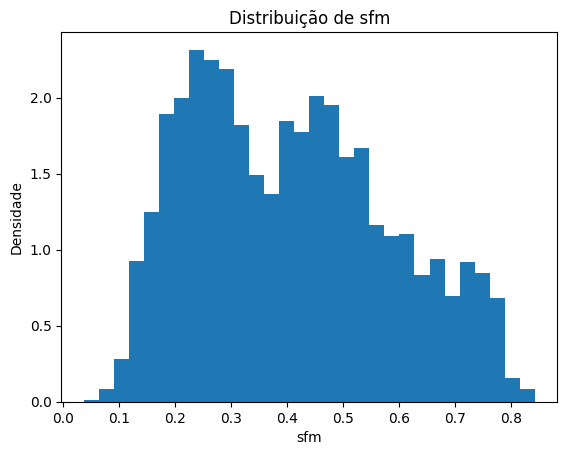
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 14)*



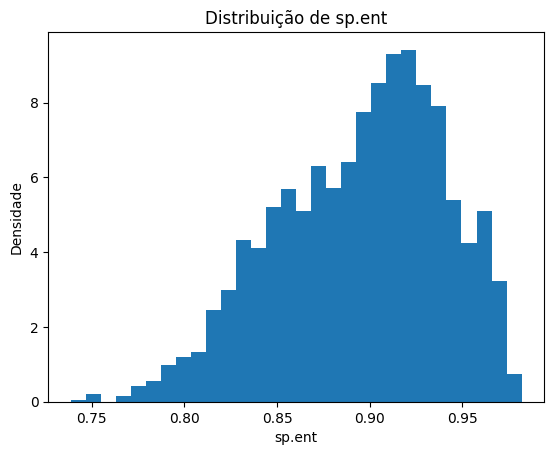
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 15)*



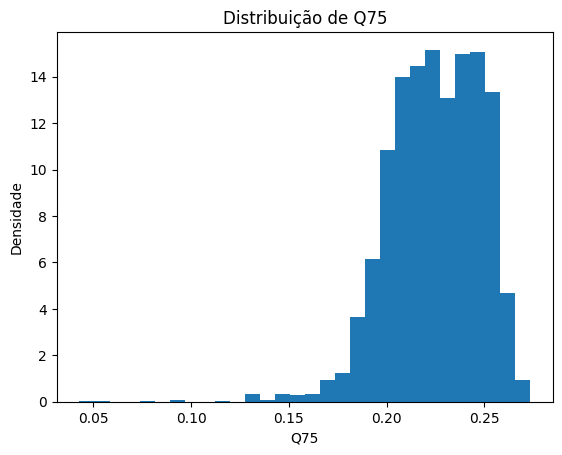
*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 16)*



*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 17)*



*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 18)*



*Figura A2 — Histogramas/densidades por feature separados por classe. (parte 19)*

## 3.3 Correlações (multicolinearidade)

Pares com |r|>0.9 detectados: meanfreq×centroid (1.000), skew×kurt (0.977), median×meanfreq (0.925), Q25×meanfreq (0.911), e pares equivalentes envolvendo centroid. Mantivemos todas as variáveis e deixamos a regularização resolver a redundância.

# 4) Parte A.2 — Seleção de hiperparâmetros (GridSearchCV)

Melhor combinação: penalty='l2', C=10.0, solver='liblinear', max\_iter=1000; acurácia média CV=0.9767.

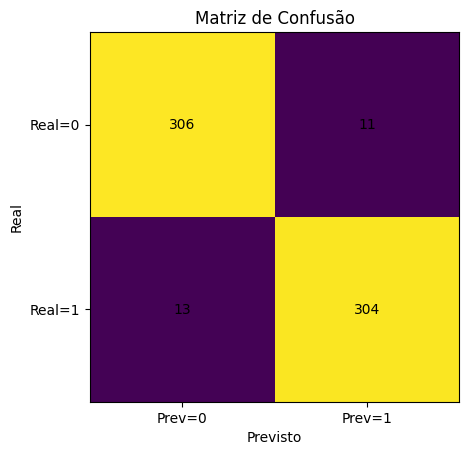
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| penalty | C | mean\_test\_score | std\_test\_score | mean\_train\_score |
| l2 | 10.00 | 0.976712 | 0.008518 | 0.977703 |
| l1 | 100.00 | 0.976317 | 0.008200 | 0.977604 |
| l2 | 100.00 | 0.976317 | 0.008200 | 0.977604 |
| l1 | 10.00 | 0.975923 | 0.007650 | 0.977604 |
| l1 | 1.00 | 0.975528 | 0.008166 | 0.977703 |
| l2 | 1.00 | 0.975133 | 0.007774 | 0.976815 |
| l1 | 0.10 | 0.973161 | 0.006810 | 0.974743 |
| l2 | 0.10 | 0.971975 | 0.009059 | 0.973954 |
| l1 | 0.01 | 0.964872 | 0.010268 | 0.964877 |
| l2 | 0.01 | 0.963686 | 0.010956 | 0.964877 |

# 5) Parte A.3 — Modelo final e avaliação no teste

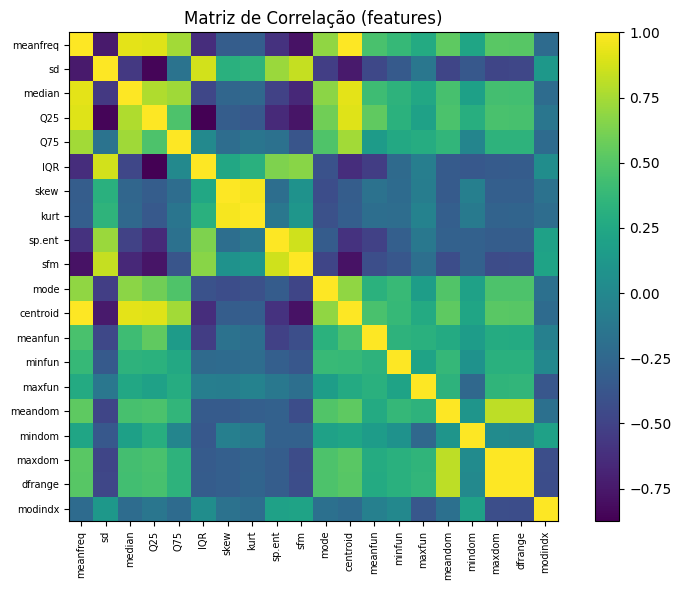
Métricas no conjunto de teste:

|  |  |
| --- | --- |
| Acurácia | 0.9621 |
| Precisão | 0.9651 |
| Recall | 0.9590 |
| F1-score | 0.9620 |

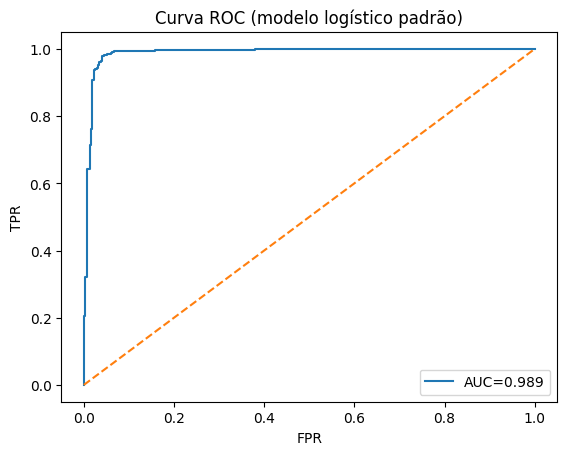
Interpretação: Precisão e Recall muito próximos indicam um classificador equilibrado — coerente com o balanceamento 50/50 das classes.



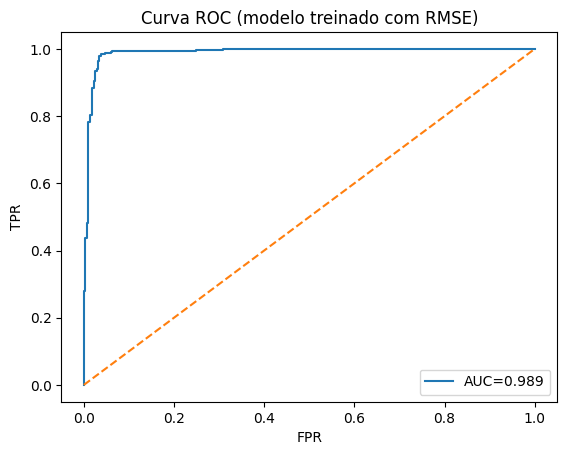
*Figura A5 — Matriz de confusão (heatmap). (parte 1)*



*Figura A5 — Matriz de confusão (heatmap). (parte 2)*



*Figura A6 — Curva ROC e AUC. (parte 1)*



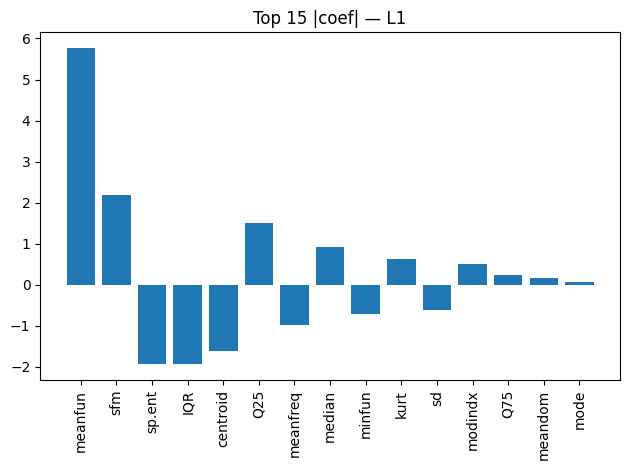
*Figura A6 — Curva ROC e AUC. (parte 2)*

# 6) Parte A.4 — Coeficientes (L1 × L2) e interpretação

Comparação direta dos coeficientes mais relevantes sob L1 e L2 (usando seus resultados):

|  |  |  |
| --- | --- | --- |
| feature | coef\_L1 | coef\_L2 |
| meanfun | 5.764898 | 5.682555 |
| sfm | 2.175116 | 2.073705 |
| sp.ent | -1.941423 | -1.832246 |
| IQR | -1.938414 | -1.765976 |
| centroid | -1.610189 | -1.181830 |
| Q25 | 1.502476 | 1.589438 |
| meanfreq | -0.974266 | -1.181830 |
| median | 0.928197 | 0.860589 |
| minfun | -0.712064 | -0.696978 |
| kurt | 0.634656 | 0.581781 |
| sd | -0.619757 | -0.550627 |
| modindx | 0.497065 | 0.491182 |
| Q75 | 0.225542 | 0.074424 |
| meandom | 0.163266 | 0.157559 |
| mode | 0.073492 | 0.070518 |

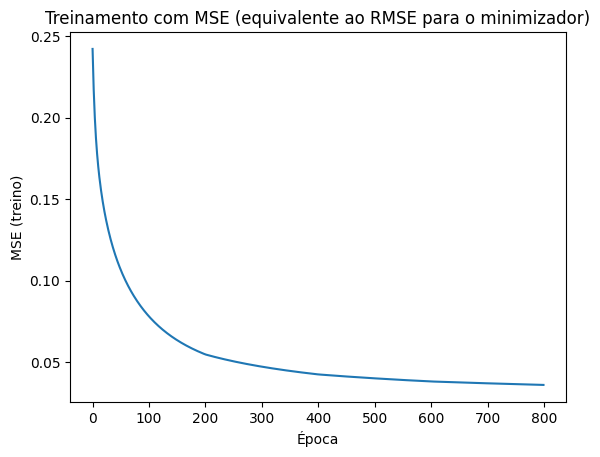
Observação: o melhor L1 (C=100.0) não zerou variáveis — penalização fraca. Para seleção dura, C menores tenderiam a induzir esparsidade. A correlação alta entre meanfreq e centroid explica a partilha de peso sob L2. Sinais indicam o efeito na log-odds de 'female' (positivos aumentam, negativos reduzem).



*Figura A8 — Barras Top |coef| (L1 ou L2).*

# 7) Parte B — Logística “do zero” com RMSE (comparativo)

Convergência do treinamento (MSE por época):



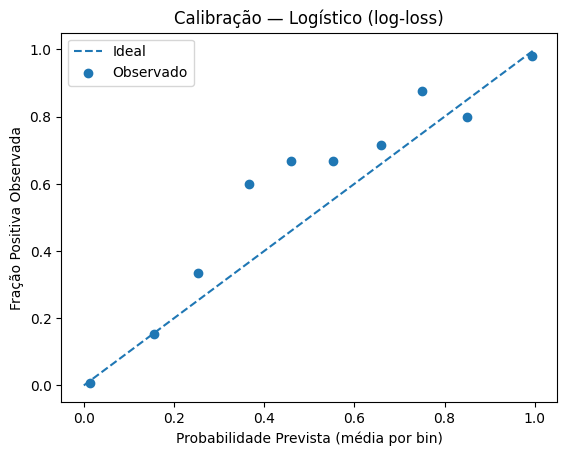
*Figura B1 — Curva de treinamento (MSE × época).*

Desempenho no teste (RMSE-model):

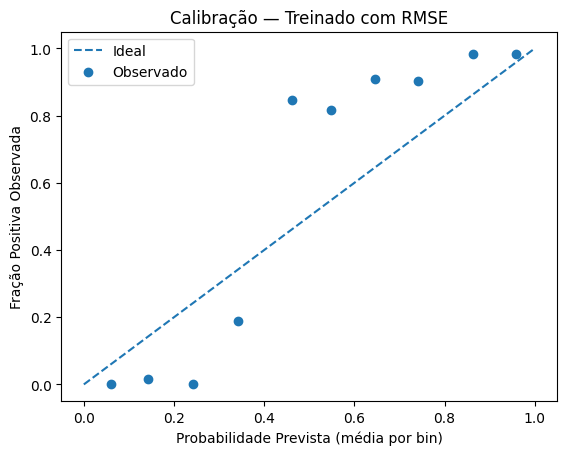
|  |  |
| --- | --- |
| Acurácia | 0.9590 |
| Precisão | 0.9678 |
| Recall | 0.9495 |
| F1-score | 0.9586 |
| RMSE (teste) | 0.2017 |

Calibração de probabilidades (quanto mais próximo da diagonal, melhor; Brier menor = melhor):

**Brier (Logístico padrão):** 0.0287 | **Brier (RMSE):** 0.0407



*Figura B2 — Calibração: modelo logístico (log-loss).*

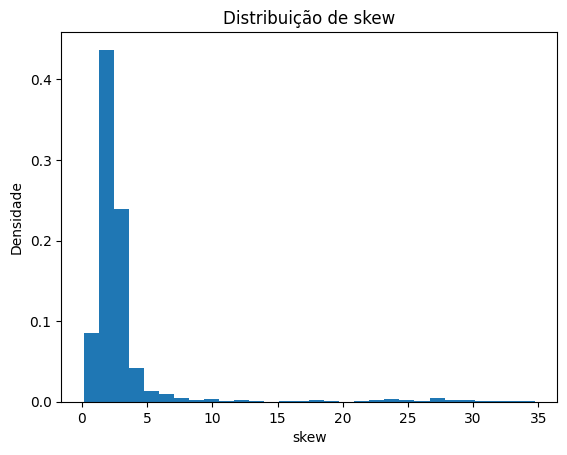


*Figura B3 — Calibração: modelo RMSE.*

# 8) Conclusões

* Modelo recomendado: Regressão Logística com L2 e C=10 (calibração superior e ótimo desempenho).
* Modelo RMSE: métricas próximas, mas calibração inferior (Brier maior).
* Para seleção de variáveis, testar L1 com C menores para induzir esparsidade.
* Multicolinearidade observada (meanfreq×centroid) tratada por L2; sinais coerentes com a EDA.

# Apêndice — Figuras adicionais



*Figura AP — Figura adicional (sem mapeamento automático).*