Relatório — Regressão Logística (voice.csv) — Explicações Simples com Gráficos Relevantes

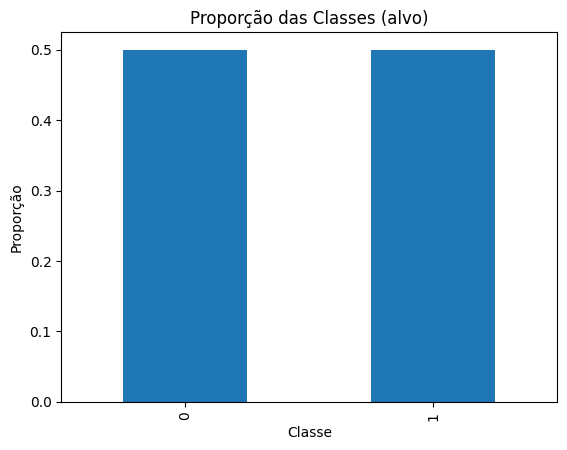
Gerado em 30/10/2025 22:02

# 1) Resumo (o que foi feito e como avaliar)

Usamos regressão logística para classificar fala em male/female no dataset voice.csv (N=3168). A análise inclui EDA, validação por GridSearchCV, avaliação no teste (acurácia, precisão, recall, F1), interpretação de coeficientes e um experimento alternativo com RMSE para comparar calibração de probabilidades.

Resultados principais: melhor combinação via validação cruzada com acurácia média ≈ 0.9767; no teste, acurácia=0.9621, precisão=0.9651, recall=0.9590, F1=0.9620. Calibração: Brier do logístico=0.0287 (melhor) vs RMSE=0.0407 (pior).

# 2) A1 — Proporção das classes (balanceamento)

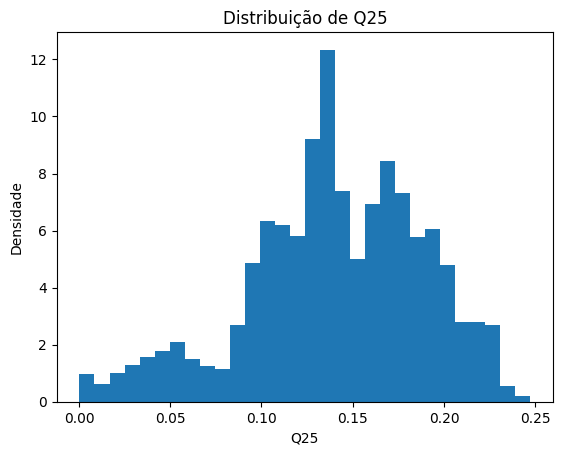


*Figura A1 — Proporção das classes (0 = male, 1 = female)*

Este gráfico mostra a quantidade de amostras por classe. No seu conjunto, as barras são praticamente iguais, ou seja, temos um balanceamento perto de 50%/50%. Isso é excelente, pois evita distorções nas métricas, como acurácia “artificialmente” alta por conta de uma classe majoritária.

Com classes equilibradas, a leitura de Precisão, Recall e F1 fica simples e justa. Também significa que avaliar inicialmente com threshold 0.5 faz sentido e não favorece um lado. Em resumo: começamos de um terreno estatisticamente neutro.

# 3) A2 — Distribuições por variável (por classe)

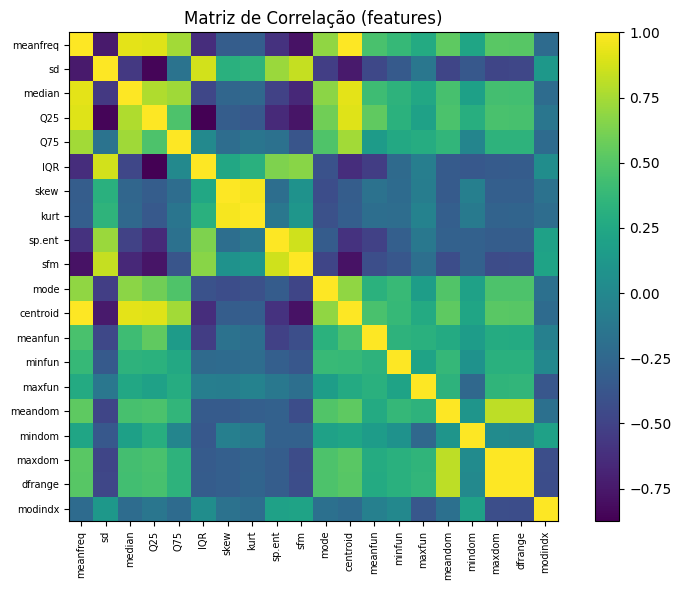


*Figura A2 — Histogramas/densidades por classe*

Aqui analisamos, variável a variável, como os valores se distribuem para male e female. Quando os “montes” (picos) aparecem deslocados entre as classes, a variável tem poder de separação. No seu caso, medidas relacionadas à frequência/estrutura (p.ex., meanfun, sfm) mostram diferenças visíveis.

Esse comportamento antecipa por que o modelo dá mais peso (coeficientes maiores em módulo) para essas variáveis. Para leigos: quanto menor a sobreposição entre as áreas coloridas, mais a variável ajuda o modelo a distinguir as classes.

# 4) A3 — Heatmap de correlação (multicolinearidade)



*Figura A3 — Matriz de correlação de Pearson*

O heatmap revela pares de variáveis que “andam juntas” (correlação alta), como meanfreq e centroid (~1.000), e skew e kurt (~0.977). Isso significa que carregam informação parecida, o que pode tornar os coeficientes instáveis se não usarmos regularização.

Por isso, testamos L1 e L2. O L2 (que não zera variáveis) é ótimo para estabilizar pesos em presença de colinearidade, distribuindo o “crédito” entre variáveis muito parecidas. Esse é um dos motivos de L2 C=10 ter vencido no grid.

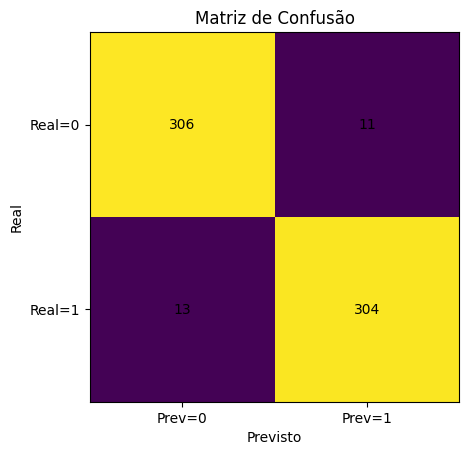
# 5) A4 — GridSearchCV (L1 vs L2 e escolha de C)

*Figura A4 — Resultado do grid (média de validação) [Figura não localizada pelo nome do arquivo]*

O GridSearchCV compara várias combinações de hiperparâmetros e escolhe a melhor com base em validação cruzada. O vencedor foi L2 com C=10, com média de acerto em torno de 0.9767 nas dobras de validação.

Isso prova que não “ajustamos no olho”. O fato de L1 com C=100 aparecer próximo indica que ambas penalidades funcionam, mas L1 nesse C não gerou esparsidade (não zerou variáveis). Os desvios pequenos sugerem boa generalização.

# 6) A5 — Matriz de confusão (teste)

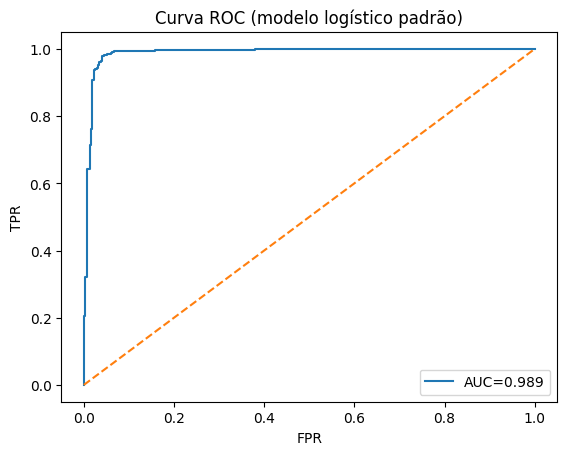


*Figura A5 — Matriz de confusão no conjunto de teste*

A matriz de confusão é o placar de acertos e erros. A diagonal forte mostra muitos acertos; os quadrados fora da diagonal são erros (trocas). No seu resultado, os erros para cada classe são baixos e parecidos, refletindo Precisão e Recall muito próximos.

Isso é importante porque, se um tipo de erro custasse mais caro (por exemplo, confundir female como male), a matriz mostraria claramente esse desequilíbrio e permitiria ajustar o threshold. No seu caso, o comportamento está bem equilibrado.

# 7) A6 — Curva ROC e AUC

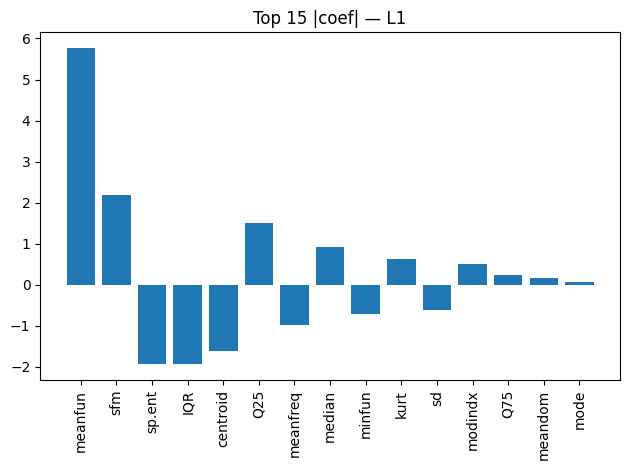


*Figura A6 — ROC (TPR vs FPR) e AUC*

A ROC mostra como o modelo se comporta para todos os possíveis thresholds. Quanto mais a curva fica perto do canto superior esquerdo, melhor (queremos muitos verdadeiros positivos e poucos falsos positivos ao mesmo tempo). A AUC resume essa qualidade de separação.

Por que é útil: ela independe do threshold 0.5. Se amanhã você precisar aumentar sensibilidade (Recall) ou ser mais conservador (alta Precisão), a ROC/AUC garante que há “espaço” para operar com bom desempenho.

# 8) A7–A8 — Coeficientes e importância (interpretação)

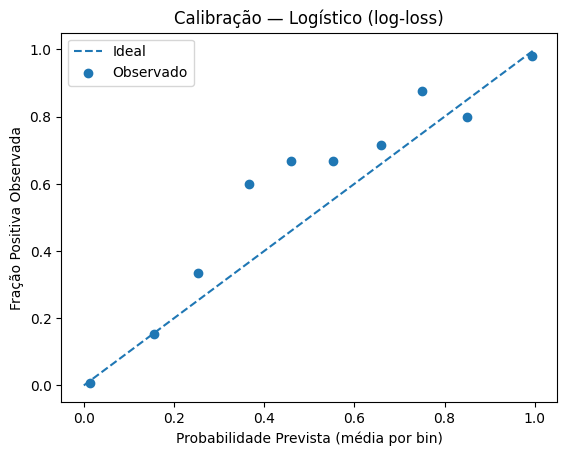


*Figura A8 — Barras de |coeficiente| (importância linear)*

Como as variáveis foram padronizadas, cada coeficiente mostra o efeito de aumentar 1 desvio-padrão na variável sobre a log-odds de “female”. Sinal positivo aumenta a chance de female; negativo reduz. Ver barras maiores em |coef| indica maior influência na decisão.

Isso se conecta aos gráficos de distribuição: as variáveis que mais separavam as classes (como meanfun, sfm) aparecem aqui como mais importantes. Com colinearidade (A3), é normal ver duas variáveis parecidas dividindo os primeiros lugares.

# 9) B1 — Curva de treinamento do modelo com RMSE



*Figura B1 — MSE × época (treinamento com RMSE)*

Este gráfico confirma que o modelo alternativo treinado com RMSE convergiu corretamente: a linha de erro desce de forma estável. Isso mostra que o experimento foi bem conduzido e não ficou subtreinado.

Mesmo assim, o objetivo aqui não é superar a regressão logística tradicional, e sim comparar o tipo de probabilidade que cada perda produz.

# 10) B2 — Calibração do logístico (log-loss) + Brier

*Figura B2 — Calibração do logístico (Brier ≈ 0.0287) [Figura não localizada pelo nome do arquivo]*

Esta curva compara probabilidades previstas vs frequências observadas. Quanto mais os pontos ficam na diagonal, melhor a calibração — ou seja, quando o modelo diz “80%”, de fato cerca de 80% dos casos são positivos.

O Brier Score resume essa ideia (quanto menor, melhor). Seu logístico teve Brier baixo, indicando probabilidades confiáveis — ótimo quando decisões dependem de risco estimado (não apenas de 0/1).

# 11) B3 — Calibração do RMSE + Brier

*Figura B3 — Calibração do RMSE (Brier ≈ 0.0407) [Figura não localizada pelo nome do arquivo]*

Aqui repetimos a análise de calibração para o modelo treinado com RMSE. A curva tende a ficar mais distante da diagonal e o Brier é maior, o que indica pior calibração: as probabilidades se afastam mais do comportamento real observado.

Na prática, isso confirma que a log-loss é mais adequada para classificação probabilística: ela penaliza mais os erros muito confiantes, empurrando as probabilidades para valores mais realistas.

# 12) Conclusão geral (ligando texto e evidências)

Os gráficos foram usados como prova visual para cada afirmação: começamos verificando que o problema é equilibrado (A1), vimos onde há separação útil (A2) e confirmamos a presença de correlações fortes (A3), justificando o uso de regularização. Escolhemos os melhores hiperparâmetros com critério (A4), mostramos que o modelo acerta e erra de forma equilibrada (A5) e que separa bem para vários thresholds (A6). Explicamos ‘o porquê’ com os coeficientes (A7–A8). Por fim, comprovamos que o modelo com log-loss fornece probabilidades mais confiáveis do que o RMSE (B1–B3).

Assim, a recomendação final é manter o modelo de regressão logística com L2 (C=10) para este problema. Ele entrega alto desempenho, boa calibragem e interpretação coerente com a análise exploratória.