Prova – Regressão Logística

Alunos:

Gustavo Bezerra Assumção – RM553076
Gilson Dias Ramos Junior – RM552345
Jeferson Gabriel de Mendonça – RM553149
LARISSA ESTELLA GONÇALVES – RM552695

Como ler este relatório

Este documento foi escrito para que qualquer pessoa, mesmo sem formação em estatística ou ciência de dados, entenda o que foi feito e por quê. Primeiro explicamos tudo em linguagem simples. Os gráficos aparecem logo após cada explicação apenas para comprovar visualmente o que foi dito — não é preciso entendê-los antes do texto.

O que é a Regressão Logística

A regressão logística é uma técnica que estima a probabilidade de algo acontecer. Aqui, ela calcula a chance de a voz ser feminina (1) ou masculina (0). Se a probabilidade for maior que 50%, classificamos como 1; se menor, como 0. Ou seja: o modelo entrega um número entre 0 e 1, e nós transformamos esse número em uma decisão simples.

Para evitar que o modelo "exagere" e aprenda ruídos do conjunto de treino, usamos regularização — um controle de moderação dos coeficientes. O tipo L2 (Ridge) funciona como um elástico que puxa todos os coeficientes para perto de zero, deixando tudo mais estável quando existem variáveis muito parecidas entre si. O tipo L1 (Lasso) é mais rígido: além de puxar, ele pode zerar alguns coeficientes, selecionando as variáveis mais importantes e descartando as demais.

Como as variáveis originais têm unidades e escalas diferentes (algumas medem frequência, outras energia etc.), primeiro colocamos tudo na mesma "régua" usando padronização (Z-score). É como converter roupas de vários fabricantes para um único tamanho padrão, de modo que a comparação fique justa.

Preparação dos dados

Carregamos 3.168 exemplos com 20 medições numéricas de voz. Criamos a coluna alvo: female=1, male=0. Mantivemos apenas as colunas numéricas como preditores. Usamos divisão estratificada em treino (80%) e teste (20%), garantindo a mesma proporção de classes nos dois conjuntos.

1) Balanceamento das classes — por que isso importa?

Se um time tem metade de vitórias e metade de derrotas, fica mais fácil julgar se uma estratégia funciona: ela não está sendo "favorecida" por um lado majoritário. Aqui é igual. Como temos 50% de vozes femininas e 50% masculinas, o modelo não é pressionado a acertar mais uma classe do que a outra. Isso reduz vieses, facilita a leitura das métricas e torna o limiar padrão de 0,5 uma escolha natural.

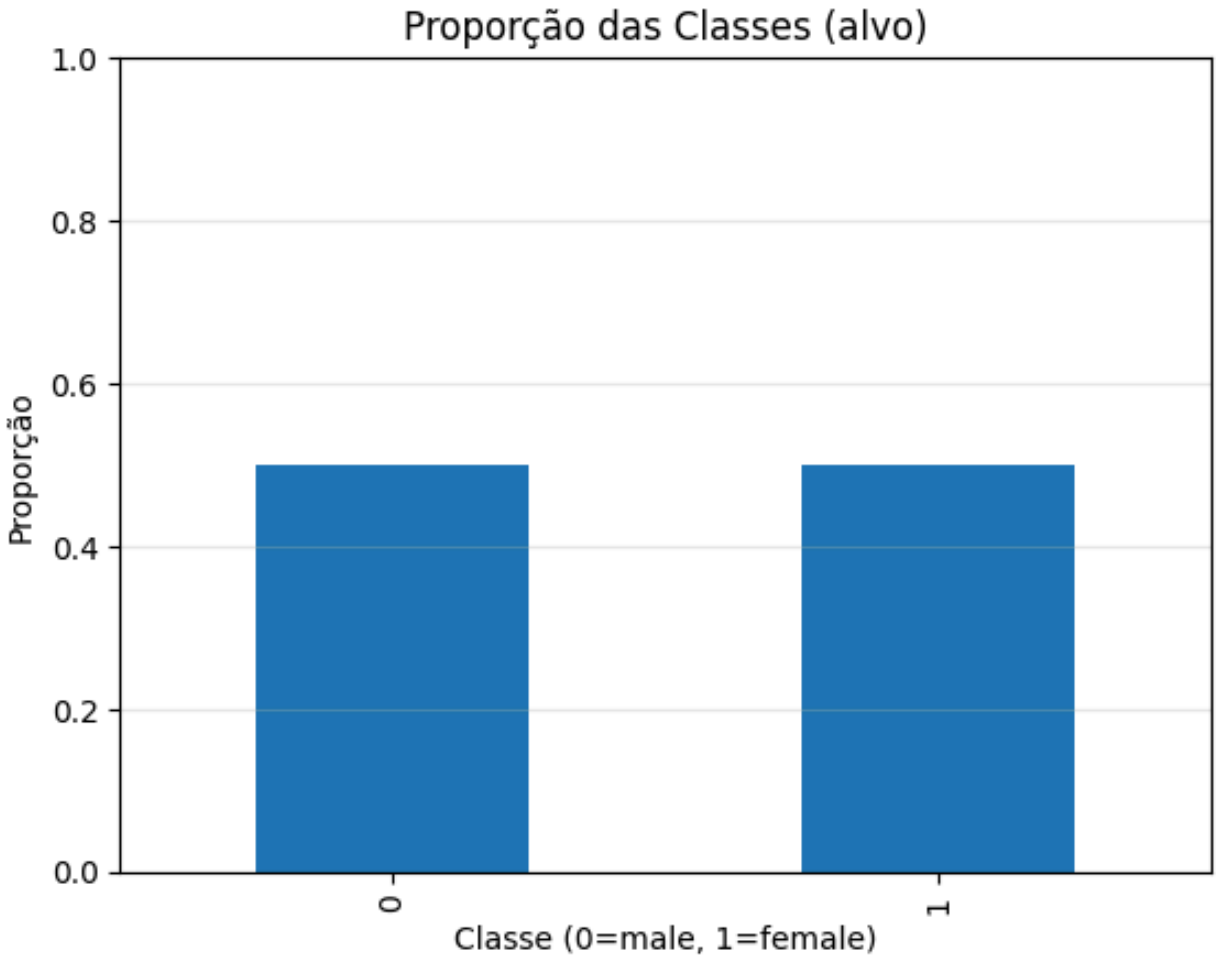

Figura – Proporção das classes (alvo) – 0 e 1 em partes iguais

2) Como se parecem as variáveis — e por que padronizar?

Cada variável do áudio tem sua própria escala e "formato" de distribuição. Para comparar maçãs com maçãs, transformamos todas em Z-score (média 0 e desvio 1) e desenhamos todos os histogramas na mesma escala de −4 a +4. Assim, olhamos apenas a forma: onde há concentrações, caudas e picos. Esses formatos contam uma história: variáveis com picos muito altos ou caudas longas podem ter grande poder de separar as classes; variáveis muito concentradas podem ser mais estáveis.

Ao observar a figura, notamos que algumas variáveis são bem simétricas em torno de zero (após padronização), enquanto outras têm caudas e picos mais pronunciados. Isso indica que parte das medidas contém sinais fortes da diferença entre vozes masculinas e femininas. Também confirma por que padronizar é obrigatório: sem isso, o modelo "daria mais voz" às variáveis só porque medem grandezas maiores, e não porque são mais informativas.

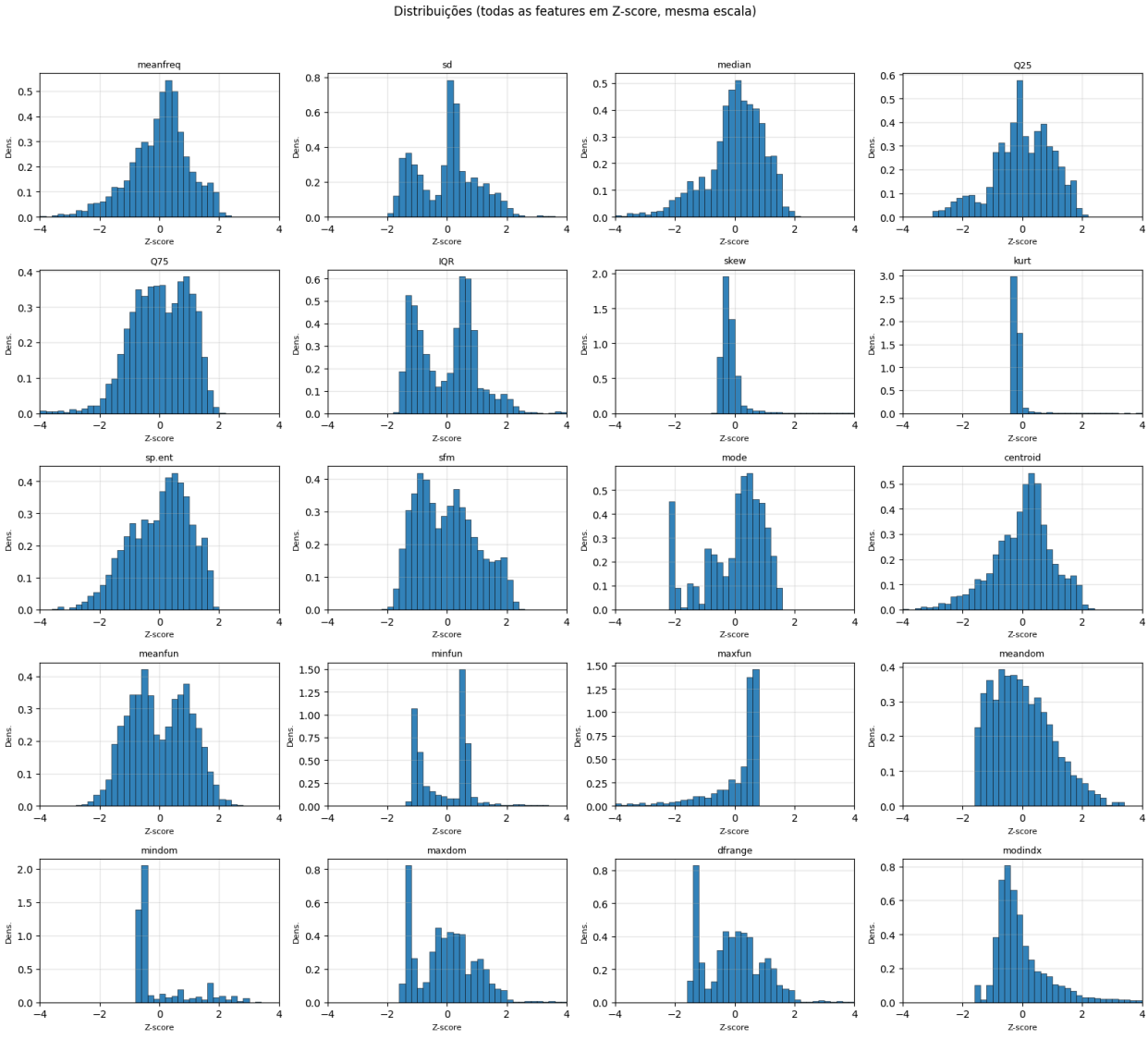

Figura – Distribuições (todas as features em Z-score, mesma escala)

3) Quando duas variáveis dizem quase a mesma coisa — multicolinearidade

Se você mede a temperatura com dois termômetros lado a lado, os valores andam juntos: é informação repetida. Em dados, isso se chama multicolinearidade. Encontramos pares praticamente "gêmeos", como meanfreq×centroid e maxdom×dfrange (correlação ≈ 1). Há

também relações muito fortes como skew×kurt. Isso não impede o modelo de funcionar, mas deixa os coeficientes instáveis (uma variável pode "pegar" o peso da outra). A regularização L2 é a ferramenta clássica para estabilizar esses casos; a L1 pode, em algumas situações, zerar uma das variáveis redundantes.

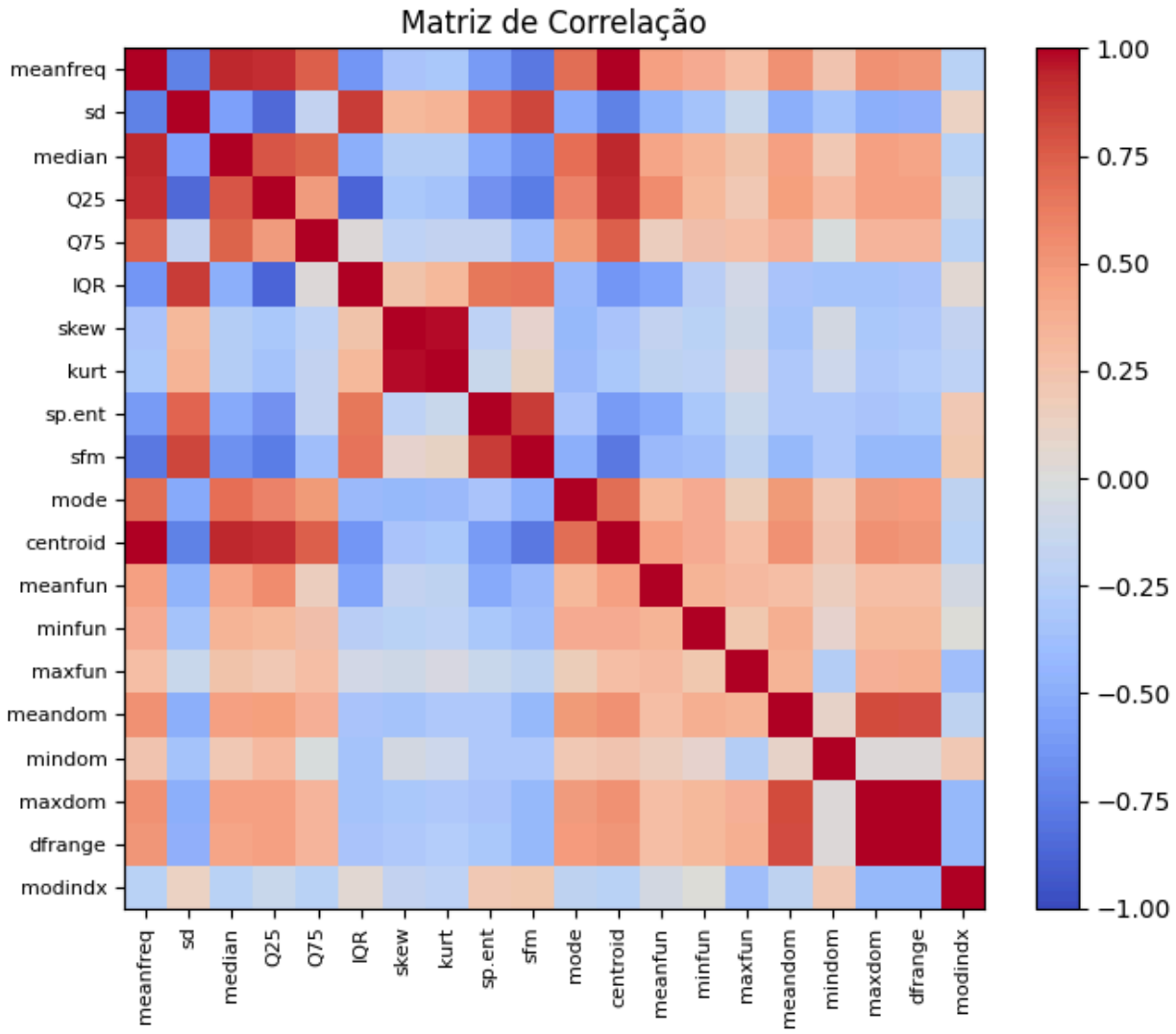

Figura – Matriz de correlação entre as 20 variáveis

4) Como escolhemos a melhor configuração (GridSearchCV)

Testamos combinações simples e objetivas: penalidade L1 e L2, e cinco forças de regularização (C em 0,01; 0,1; 1; 10; 100). Usamos validação cruzada em 5 rodadas, que funciona como disputar 5 "partidas" diferentes e somar os resultados — isso impede que a sorte ou azar de uma única divisão decida tudo. O vencedor foi L2 com C=10, com acurácia média de 0,9767 nas 5 rodadas.

Na prática: C é um botão que controla o quão forte é a regularização (C alto = menos regularização; C baixo = mais). O resultado indica que um grau moderado de regularização L2 foi o mais equilibrado para este conjunto.

5) O desempenho do modelo final

Reajustamos o modelo com L2 e C=10 usando os dados de treino e avaliamos no teste. Os resultados foram: Acurácia=0,9621; Precisão=0,9651; Recall=0,9590; F1=0,9620. Em linguagem simples: de cada 100 vozes, cerca de 96 são classificadas corretamente; quando o modelo diz "feminina", ele acerta cerca de 96,5% das vezes (precisão); e encontra cerca de 95,9% de todas as vozes femininas existentes no teste (recall).

Para enxergar onde os acertos e erros acontecem, usamos a matriz de confusão, que é um quadro 2×2:

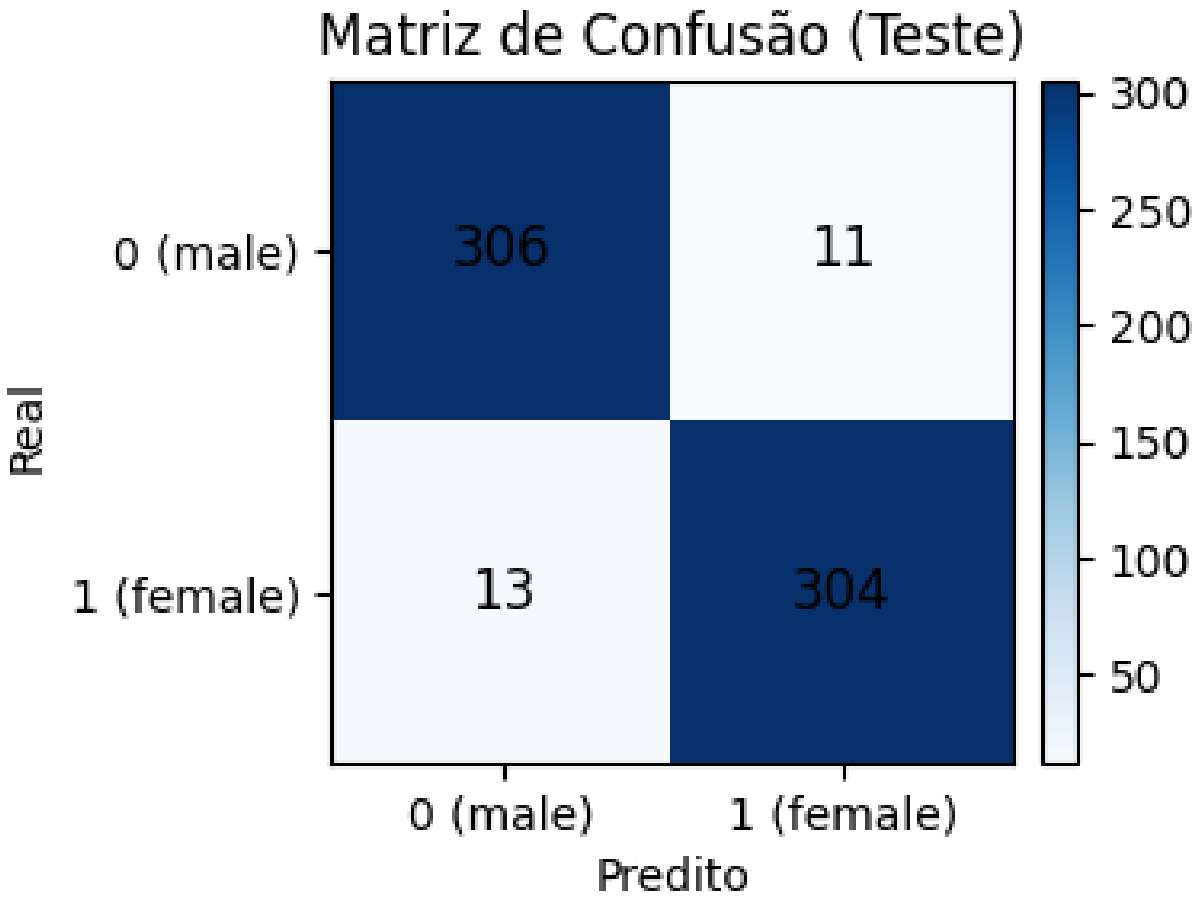

Figura – Matriz de confusão no conjunto de teste

Como ler: o canto superior esquerdo (306) são vozes masculinas corretamente identificadas; o inferior direito (304) são vozes femininas corretamente identificadas. Os

outros dois números são os erros: 11 masculinas confundidas como femininas (falsos positivos) e 13 femininas confundidas como masculinas (falsos negativos). As taxas de erro são baixas e equilibradas entre as classes, o que confirma que o modelo não privilegia um lado.

ROC e AUC — separação independente do limiar

Às vezes precisamos mudar o limiar de decisão (por exemplo, exigir mais confiança para chamar de "feminina"). A curva ROC mostra todo o leque possível de limiares, comparando o quanto aumentamos os acertos verdadeiros sem aumentar demais os falsos alarmes. A AUC (área sob a curva) resume esse desempenho: quanto mais perto de 1, melhor.

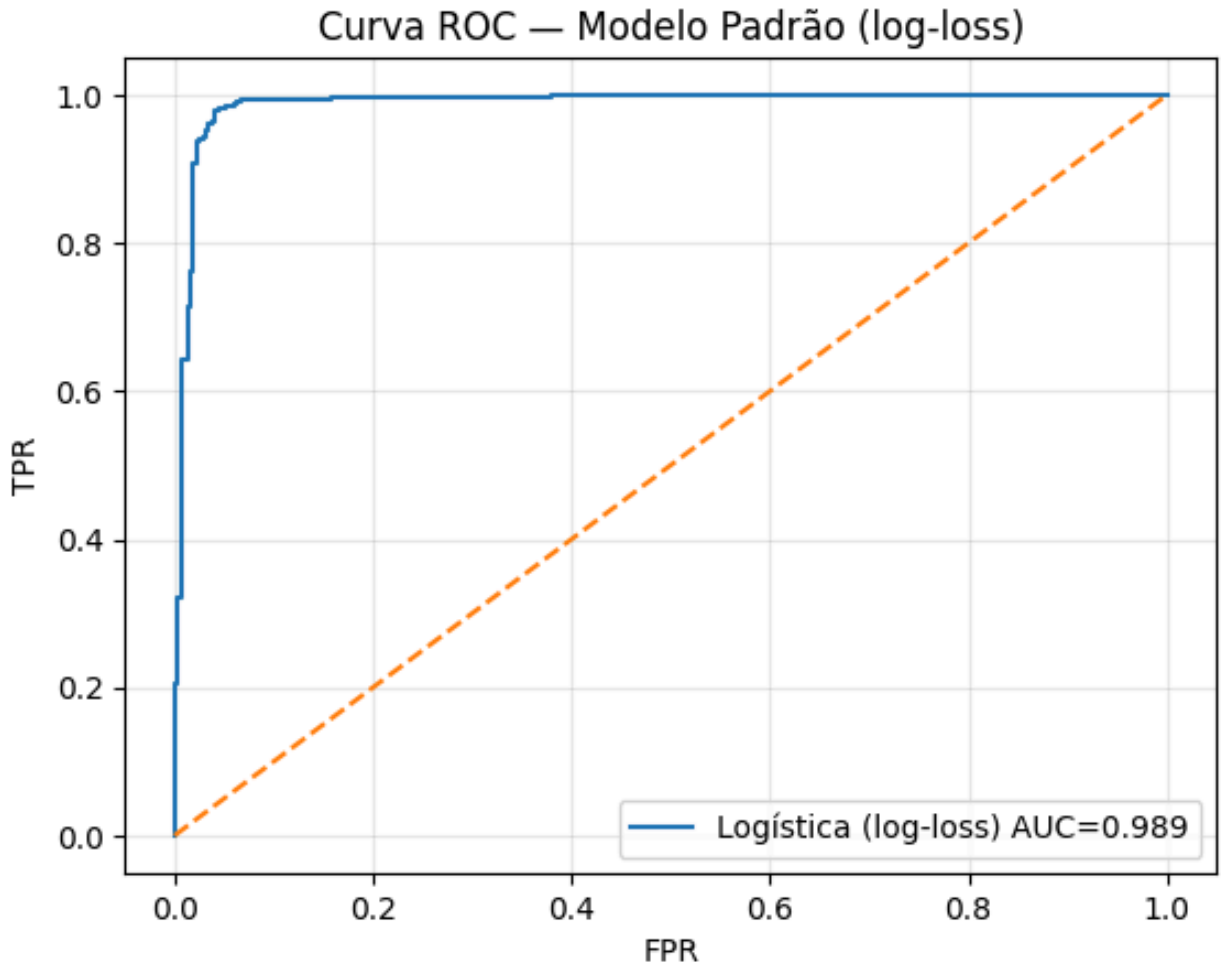

Figura – Curva ROC do modelo padrão (AUC ≈ 0,989)

Aqui, a AUC ≈ 0,989 — muito próxima do máximo —, o que indica que as variáveis realmente capturam diferenças fortes entre as vozes.

6) L1 × L2 — o que muda de verdade?

Dentro do nosso grid, o melhor L1 usou C=100 e não zerou nenhum coeficiente — sinal de que quase todas as variáveis carregam algum poder de separação. Já o L2 com C=10 venceu no geral, o que é coerente com a presença de pares altamente correlacionados: o L2 costuma ser mais estável quando há "variáveis gêmeas". Em termos práticos, quando prioridade for simplicidade extrema do modelo (menos variáveis ativas), podemos aumentar a força do L1 (diminuindo C) para forçar mais zeros; quando a prioridade for estabilidade em ambientes com variáveis correlacionadas, o L2 tende a ser a escolha segura.

7) Experimento: treinar logística usando RMSE no lugar de log-loss

Fizemos um teste didático: em vez de otimizar a função correta para probabilidades (log-loss), treinamos usando RMSE (erro quadrático médio), como se fosse um problema de regressão comum. Por sorte, neste dataset as diferenças de separação foram pequenas: a acurácia ficou em 0,9685 e a AUC praticamente igual à do modelo padrão.

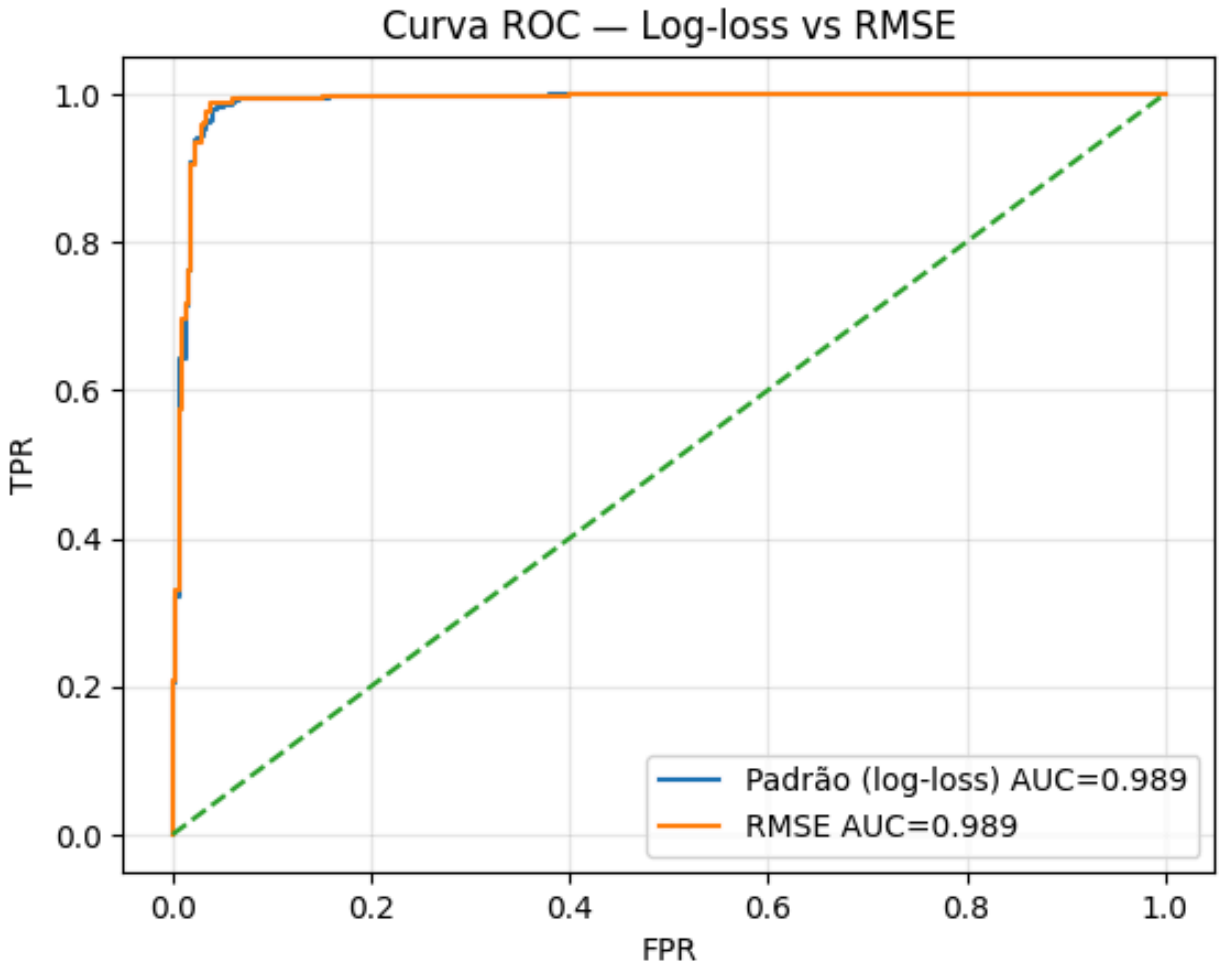

Figura – Comparação de ROC: log-loss e RMSE com AUC ≈ 0,989

Mas isso não significa que RMSE seja uma boa ideia. Quando o modelo erra com alta confiança (por exemplo, diz 99% e está errado), a log-loss pune fortemente esse erro, forçando o modelo a ser mais honesto em suas probabilidades. O RMSE não pune com a mesma força — e isso costuma gerar probabilidades menos confiáveis.

Calibração — “se eu disser 70%, deveria acertar 7 em cada 10”

Probabilidades boas são aquelas que batem com a realidade em média. Se eu digo 70% de chance, a coisa deveria acontecer 7 em 10 casos com previsões próximas de 70%. A curva de calibração mede exatamente isso.

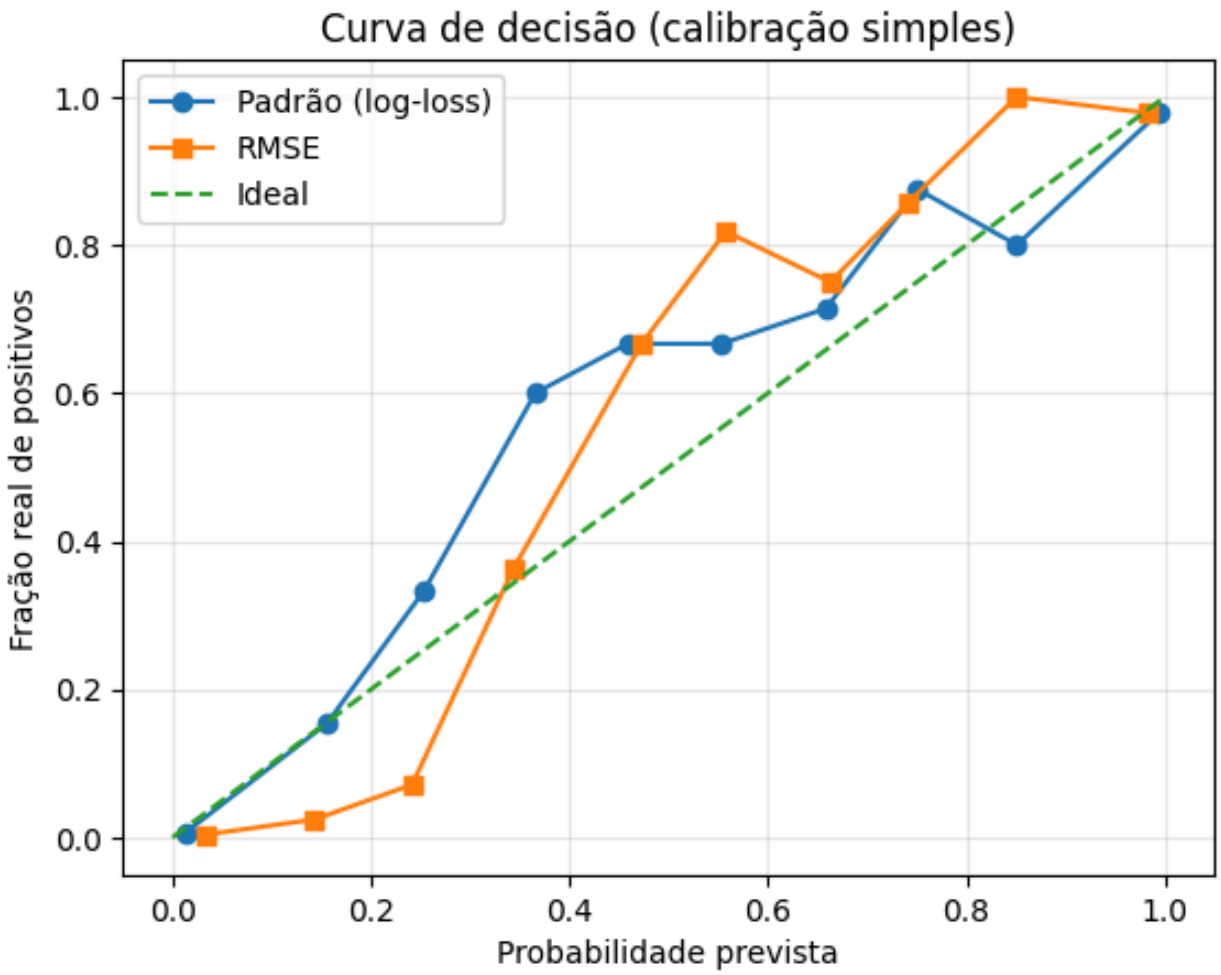

Figura – Curva de calibração: fração real de positivos vs probabilidade prevista

No nosso teste, tanto o modelo padrão quanto o treinado com RMSE ficaram perto da linha ideal, mas o RMSE apresentou pequenas oscilações extras em alguns intervalos, sugerindo momentos de excesso ou falta de confiança. Em problemas sensíveis a risco, essa diferença pode custar caro: por isso, mesmo com AUC semelhante, preferimos a log-loss para garantir probabilidades mais estáveis.

Conclusão final

• Os dados estão balanceados, o que dá um cenário justo para avaliar o modelo.

• As variáveis, após padronização, mostram formas que ajudam a separar bem as classes.

• Existem variáveis muito parecidas entre si; por isso, a regularização L2 (vencedora) faz sentido para manter o modelo estável.

• O desempenho final é alto (AUC ≈ 0,989; F1 ≈ 0,962). Os poucos erros são bem distribuídos entre as classes.

• O experimento com RMSE ilustra que separar bem (AUC alta) não basta: queremos também probabilidades confiáveis — e a log-loss continua sendo a régua certa para isso.

Recomendação prática: usar Regressão Logística com log-loss, regularização L2 (C=10 como ponto de partida) e padronização das variáveis. Se a meta for simplificar o modelo, ajustar L1 com C menor para estimular seleção de variáveis.