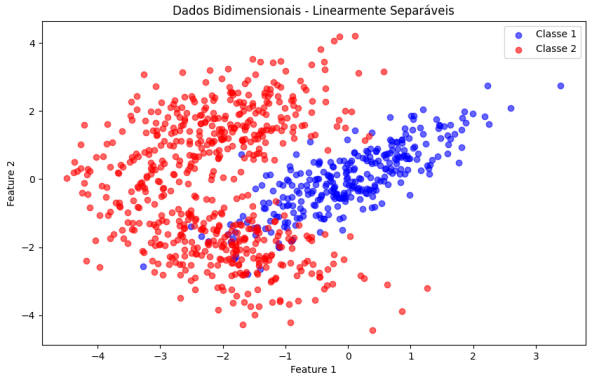
LDA e QDA

Parte 1: Gerando os Dados

Para ser objetivo, nessa parte foi feita apenas a geração do conjunto de dados que vamos utilizar no exercício.

A matriz de covariância das classes 1 e 2 são simétricas, contendo, na diagonal principal, a variância das p variáveis e, nos demais elementos, as covariâncias das variáveis.

Abaixo se encontra o gráfico do nosso dataset:



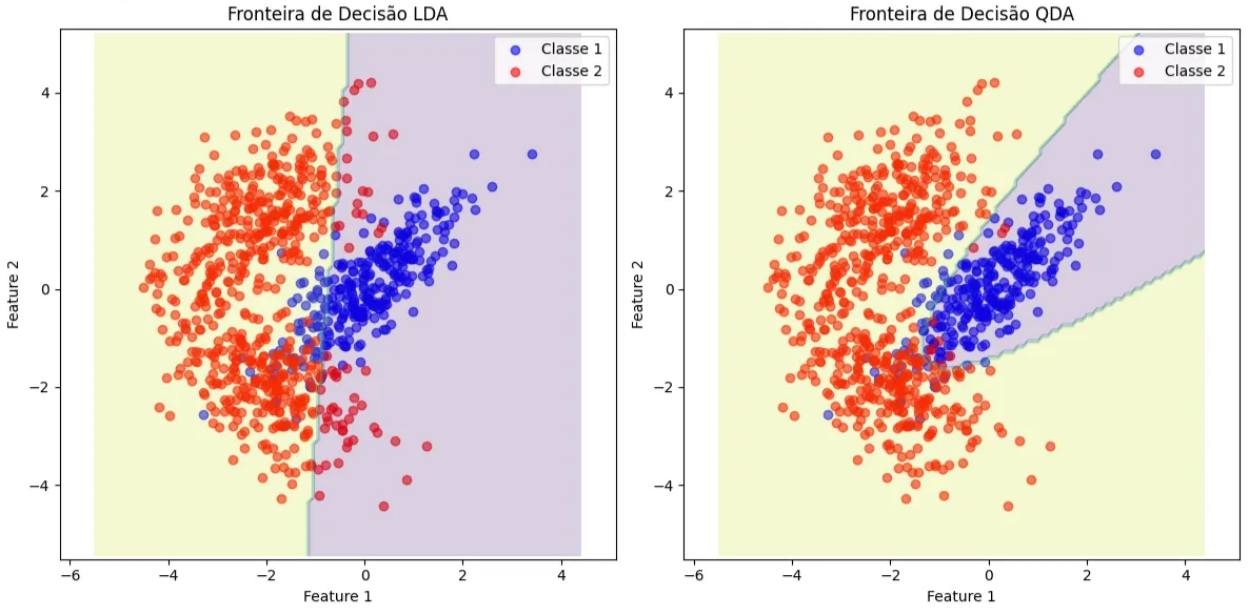
Parte 2: Treinamento e Avaliação dos Classificadores

Separamos nosso conjuntos de dados em treinamento (70%) e teste (30%) para avaliar os modelos de classificação LDA e QDA no nosso conjunto de teste.

Obtivemos as seguintes medidas estatísticas do modelo:

```
Acurácia LDA: 0.84
Precisão LDA: 0.89
Acurácia QDA: 0.94
Precisão QDA: 0.92
```

Abaixo é mostrado as fronteiras de decisões de cada modelo para efeitos de comparação e análise.



A partir do que foi mostrado, vemos que o **QDA apresentou medidas estatísticas melhores** (precisão e acurácia) em relação ao LDA, comportando melhor ao problema apresentado. Em complementar a isso, vemos que a **fronteira de decisão** do problema não é linearmente separável (por uma reta, por exemplo), o que prejudica o desempenho do LDA, que se comporta melhor para problemas linearmente separáveis.

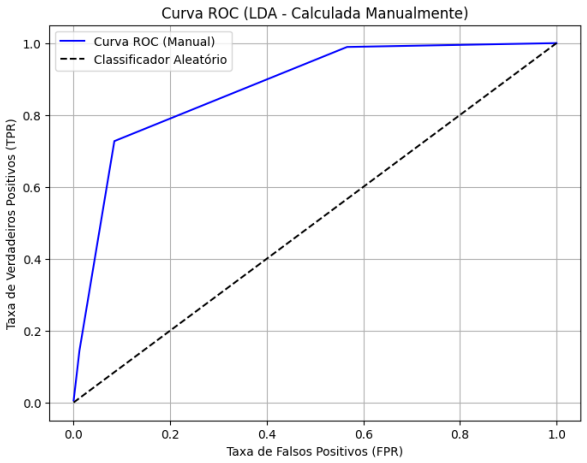
O QDA é uma técnica que surge a partir do LDA, que permite a separação não-linear dos dados, que é justamente o que queremos no nosso dataset.

Parte 3: Análise e Discussão

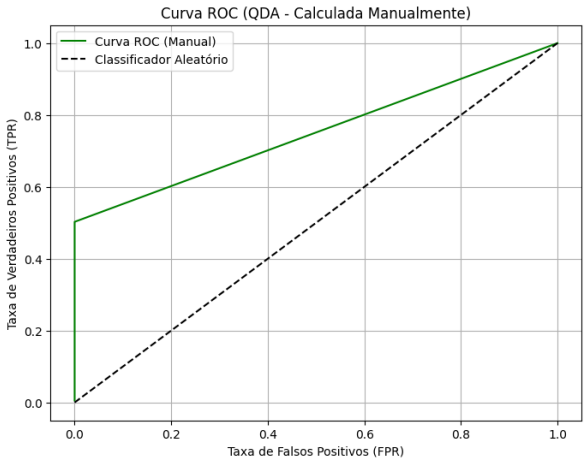
Nessa etapa vamos analisar a curva ROC (Característica de Operação do Receptor) para cada modelo. Falando de uma forma mais simples e objetiva, essa curva nos ajuda a analisar o quão bom nosso modelo está, de acordo com nossa **TPR** (Taxa de Verdadeiros Positivos) e **FPR** (Taxa de Falsos Positivos).

Vamos variar o limiar de decisão para avaliar qual se encaixa melhor no nosso modelo.

Selecionando 5 amostras do limiar de decisão entre o valor mínimo da função discriminante e o valor máximo da função discriminante:

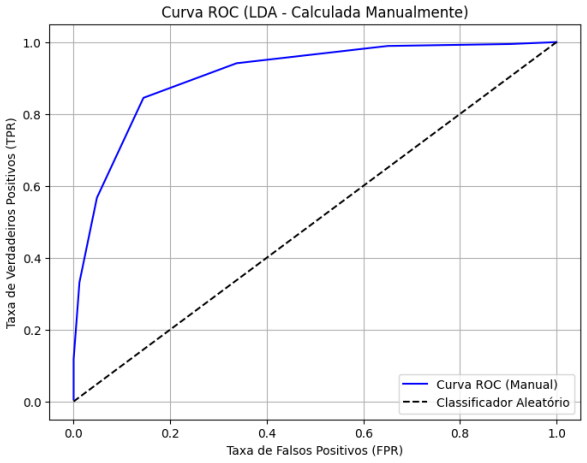


Curva ROC - LDA (Figura 1).

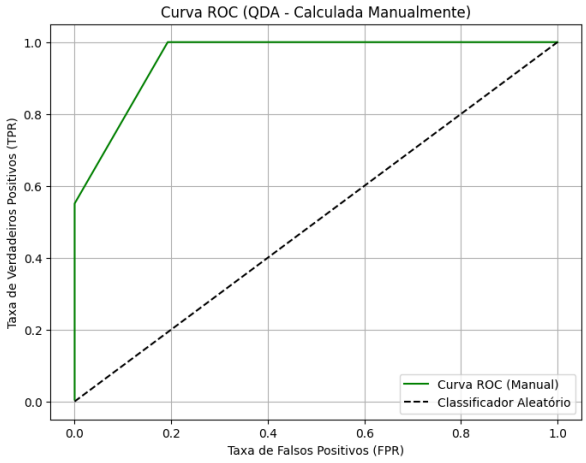


Curva ROC - QDA (Figura 2)

Selecionando 10 amostras do limiar de decisão entre o valor mínimo da função discriminante e o valor máximo da função discriminante:

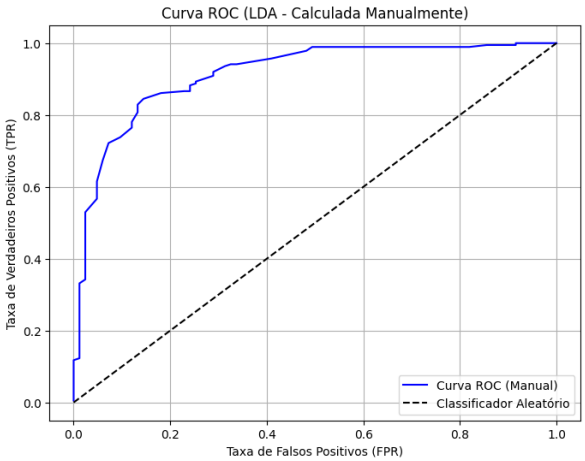


Curva ROC - LDA (Figura 3)

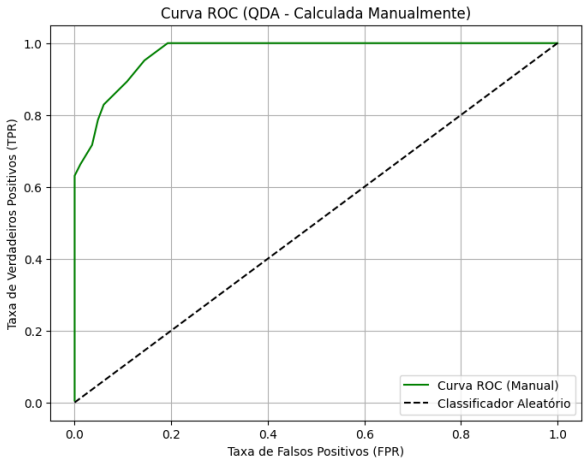


Curva ROC - QDA (Figura 4)

Selecionando 100 amostras do limiar de decisão entre o valor mínimo da função discriminante e o valor máximo da função discriminante:



Curva ROC - LDA (Figura 5)



Curva ROC - QDA (Figura 6)

Podemos continuar aumentando a quantidade de amostras para o limiar de decisão para obtermos valores ainda mais precisos se quisermos, mas com o que foi mostrado já é possível tirarmos uma conclusão a respeito da curva ROC.

Ao selecionarmos 5 amostras (Figura 1), é intuitivo pensarmos que o modelo LDA se comportou melhor em relação ao QDA devido ao fato da curva ROC se aproximar mais rapidamente a uma

$TPR = 1$, o que corrobora com nosso argumento.

Ao analisarmos o gráfico de 10 amostras (Figura 4), vemos que o comportamento do gráfico do QDA muda drasticamente, mostrando que o modelo se aproxima de uma taxa $TPR = 1$ antes do LDA, o que nos leva a fazer um outro teste com mais amostras do limiar de decisão para confirmar qual modelo possui uma melhor curva ROC.

Ao analisarmos o gráfico de 100 amostras (Figura 6), é perceptível que o QDA se mostra superior ao LDA, devido ao fato de sua curva ROC mostrar que o modelo se aproxima a um $TPR = 1$ a uma taxa de Falsos Positivos (FPR) menor que o modelo LDA, mostrando que possui uma Taxa de Falsos Positivos menor, e com

isso se mostrando um modelo mais eficaz para o problema apresentado no exercício.

Perguntas finais

Compare as precisões dos classificadores LDA e QDA no conjunto de teste e compare as
curvas ROC (plote as curvas num mesmo gráfico para comparação). Responda às perguntas:

1. Qual dos dois classificadores teve melhor desempenho em termos de precisão?

- O classificador QDA. A partir do que foi mostrado, foi visto que o QDA apresentou uma métrica de precisão maior e pela curva ROC também podemos concluir isso.

2. Com base nas fronteiras de decisão, qual parece melhor ajustado aos dados? Por quê?

- O classificador QDA. O QDA está mais ajustado aos dados devido ao fato de apresentar uma boa separabilidade entre as classes, que envolve uma separação não-linear para o problema apresentado, mas corre o risco de sofrer overfitting se o modelo ficar muito ajustado aos dados e não conseguir generalizar para possíveis novos dados que o modelo encontre.

**3. Discuta em que tipos de problemas (além deste) seria vantajoso utilizar LDA em vez de
QDA e vice-versa, considerando as suposições e diferenças entre os dois métodos.**

LDA deve ser usado, preferencialmente, quando temos uma boa separação linear entre as variáveis/classes do nosso problema.

Problemas que seria vantajoso o uso de LDA:

- Previsão de categorias de clientes: o LDA pode ser usado para classificar clientes em diferentes grupos com base no comportamento de compra ou

demografia, o que é útil para campanhas de marketing direcionadas.

- Diagnóstico de falhas: na fabricação, o LDA pode ajudar a classificar máquinas ou peças em categorias "normais" ou "defeituosas" com base em várias leituras de sensores, auxiliando na manutenção preditiva.

O QDA, por sua vez, se comporta bem aos dados quando queremos capturar uma separação não-linear do problema. O seu melhor uso é em cenários em que as classes apresentam diferentes variâncias e/ou a condição de contorno do problema é não-linear.

Problemas que seria vantajoso o uso de QDA:

- Detecção de fraudes em finanças: o QDA pode ser usado eficientemente no setor de finanças para detectar padrões de transações fraudulentas com base em dados históricos. A relação não-linear entre os diferentes indicadores financeiros pode ajudar a distinguir entre atividades financeiras regulares e fraudulentas.
- Classificação biológica: na Biologia, QDA pode ser usado para classificar espécies/genótipos com base em features morfológicas/genéticas, especialmente quando a relação entre as features e classes são não-lineares e complexas.

4. Mude o centro da classe 1 para (3,3), refaça os gráficos anteriores e analise o que aconteceu e tire suas conclusões.

Ao mudarmos o centro da classe 1, obtemos o seguinte dataset ajustado:

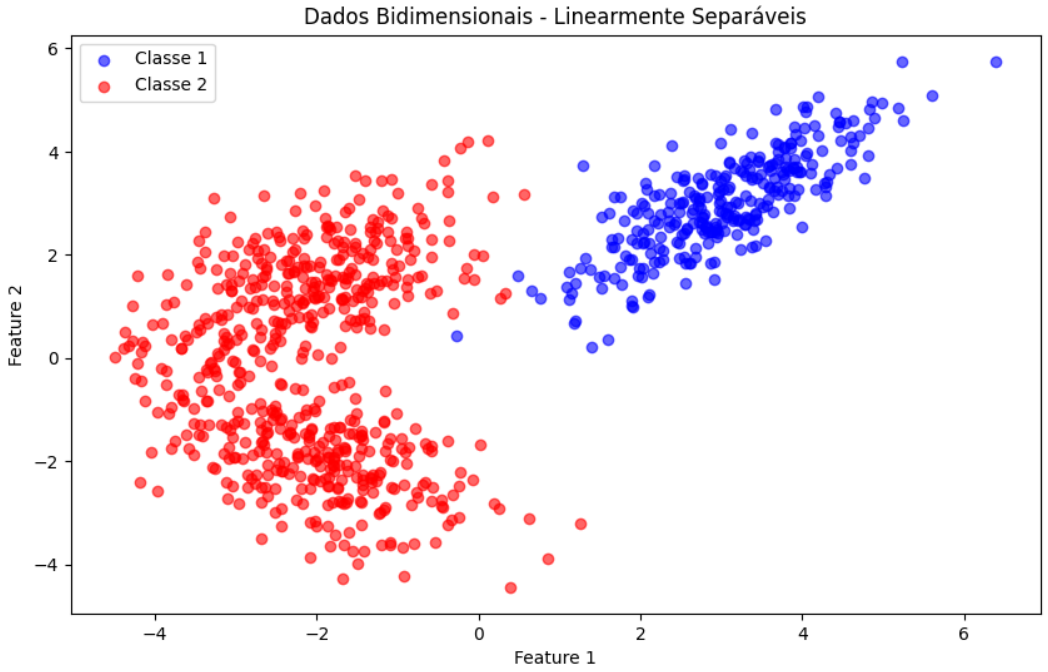


Figura 7.

Vamos treinar ambos os modelo e plotar a fronteira de decisão:

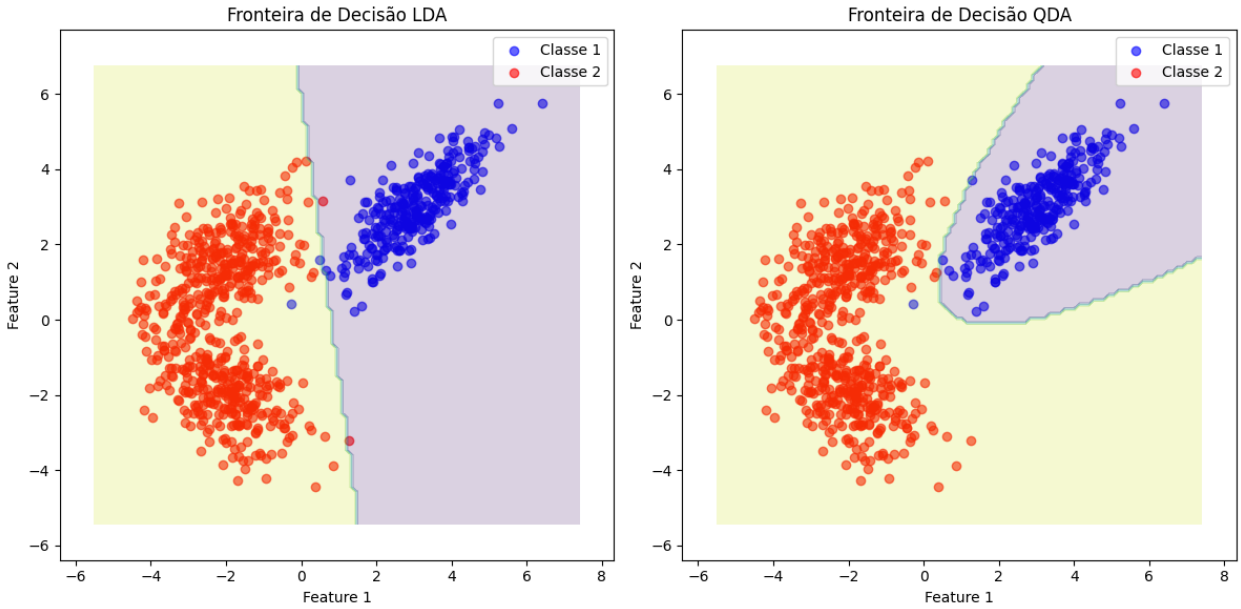


Figura 8.

A partir das fronteiras de decisão, podemos imaginar que a precisão e a acurácia de ambos os modelos ficaram bem parecidas, que se confirma quando calculamos essas métricas:

```
Acurácia LDA: 0.99
Precisão LDA: 0.99
Acurácia QDA: 1.00
Precisão QDA: 0.99
```

O QDA apresentou uma acurácia um pouco maior, porém ele corre o risco de apresentar um overfitting quando for generalizar para novos dados que possam entrar para o modelo.

Abaixo vamos analisar as curvas ROC de cada modelo:

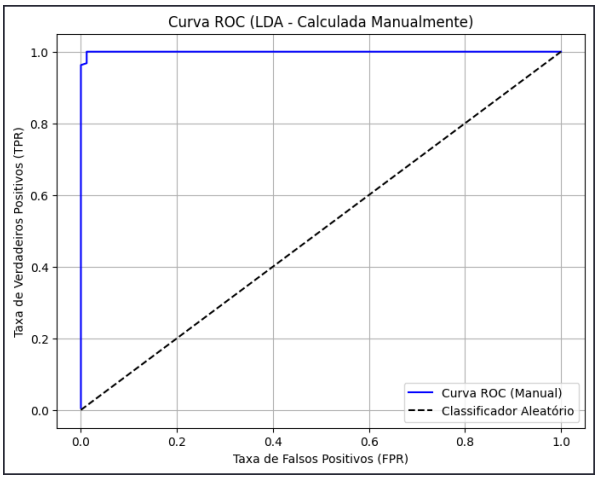


Figura 9.

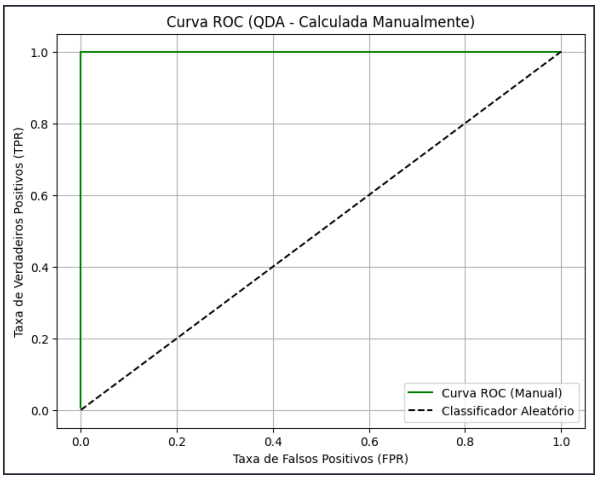


Figura 10.

A partir das curvas ROC vemos que o modelo QDA obteve o melhor resultado possível de uma curva ROC, com uma
$TPR = 1$ e uma $FPR = 0$.

Porém, como o resultado está bem parecido com a curva ROC do LDA, acredito que, nesse caso, o modelo LDA seria mais vantajoso para evitar um overfitting em relação a
possíveis novos dados que o modelo possa enfrentar, apesar do modelo QDA

mostrar um desempenho ligeiramente acima devido a sua capacidade de separação não-linear dos dados.