Machine Learning

Assignment 1

Miguel Tavares

Questão 1 – Análise dos dados

- Elevado número de valores não reais

- Glucose, BMI e Age como principais candidatas

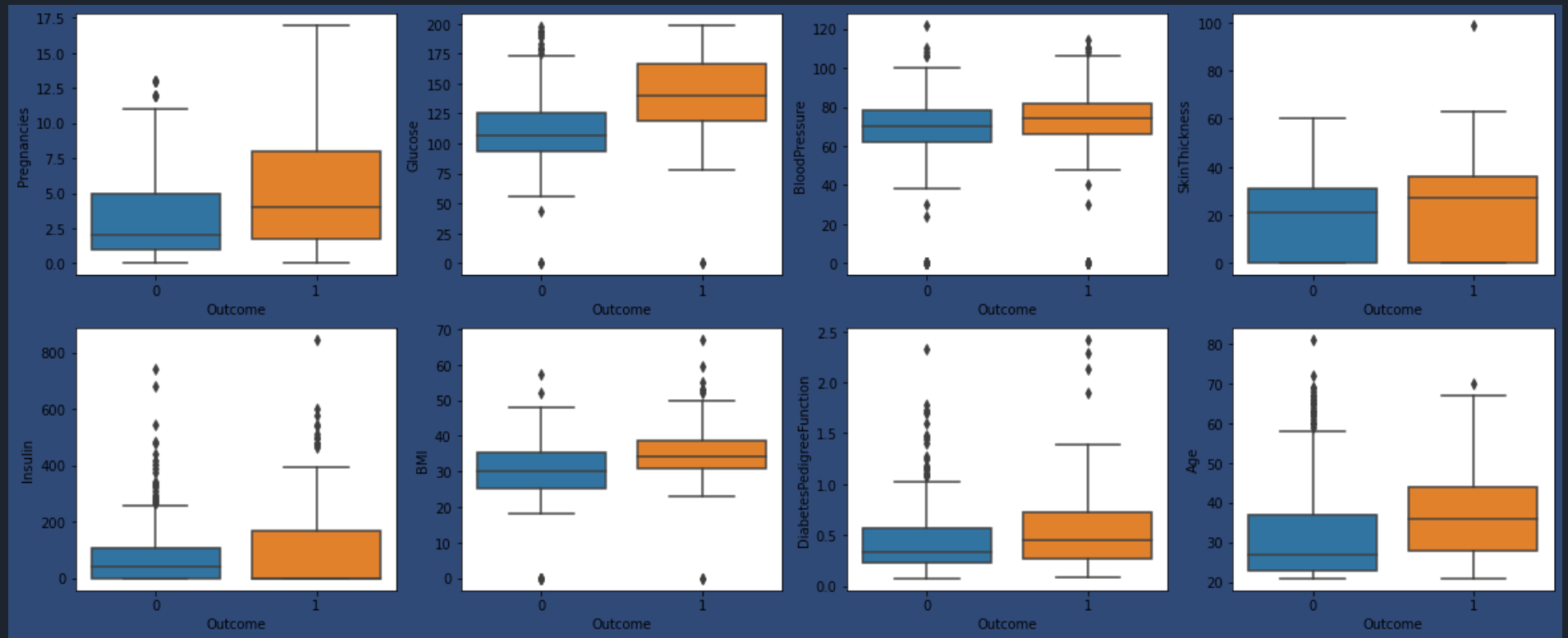

Figura 1. Box plots para as variáveis preditivas de acordo com a classe da variável alvo.

Questão 1 – Escolha de variáveis

- Qualquer umas das três variáveis separa bem as classe alvo

- Escolha da *Glucose* e *BMI* devido ao número de *outliers* da variável *Age*

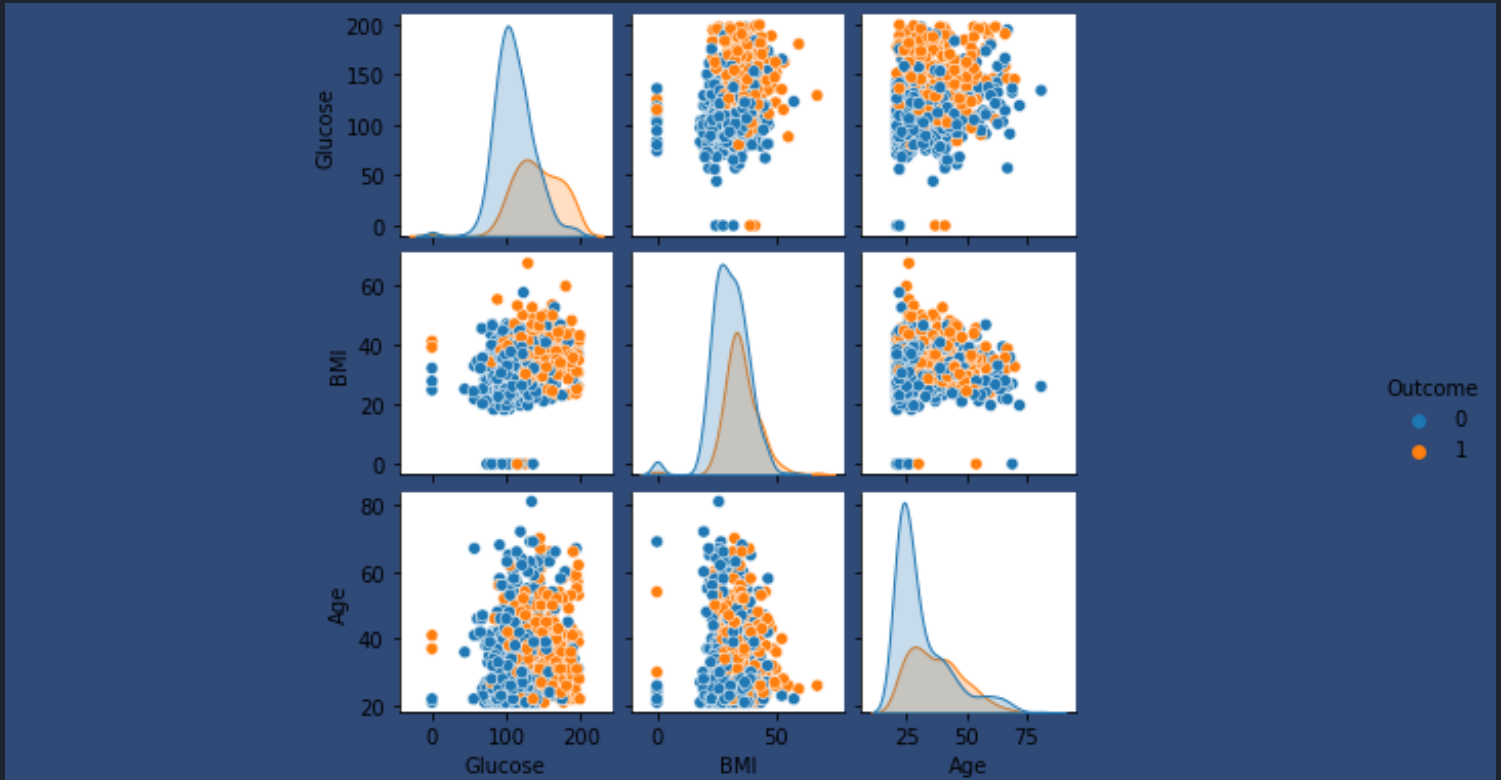

Figura 2. Gráficos de pares para três das variáveis preditivas (glucose, BMI, age) em função das variável alvo

Questão 1 – Regressão Logística e Análise Discriminante Quadrática

	Regressão Logística	Análise Discriminante Quadrática	kNN (k = 71)
F1 score on 5-fold	0,6339	0,6278	0,6111
F1 score on training set	0,5892	0,6137	0,6154
F1 score on test set	0,5581	0,5843	0,5287
Accuracy score on 5-fold	0,8015	0,7951	0,788
Accuracy score on training set	0,7587	0,7654	0,7671
Accuracy score on test set	0,7483	0,755	0,7285

Questão 1 - kNN

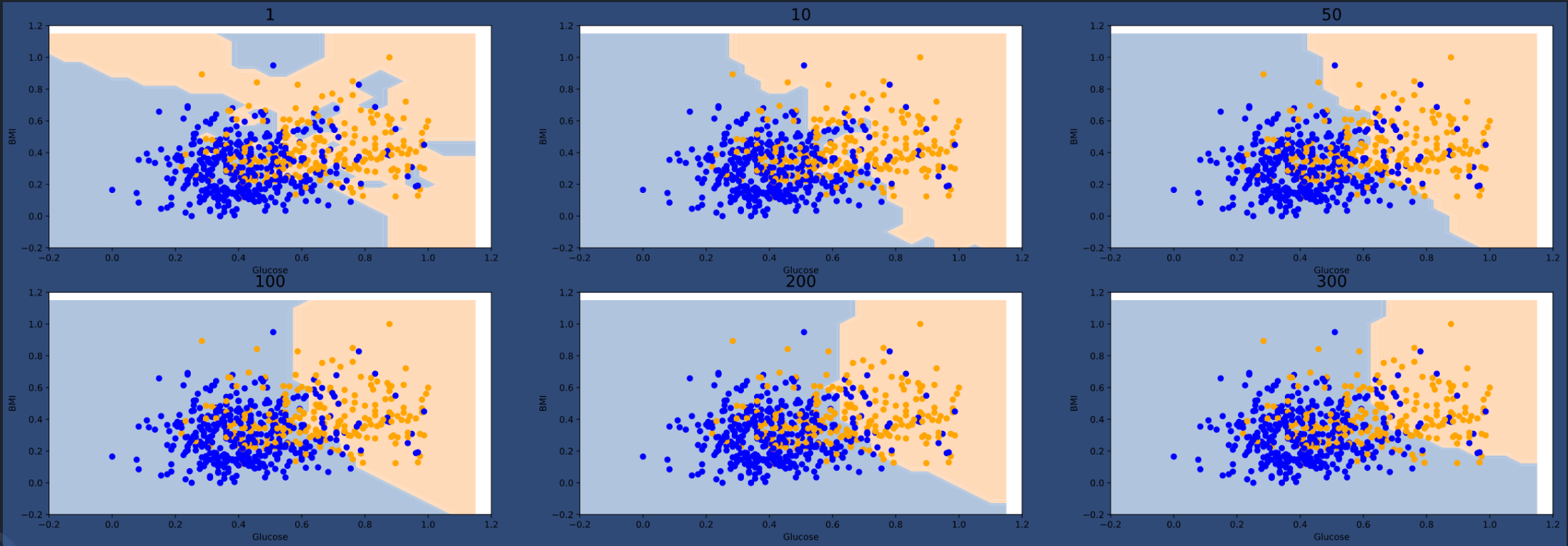

Figura 3. Fronteiras de decisão obtidas com o modelo 'k Nearest Neighbours' e usando diferentes valores de k: 1, 10, 50, 100, 200 e 300

Questão 1 – kNN (melhor k)

- Melhor k obtido: 71

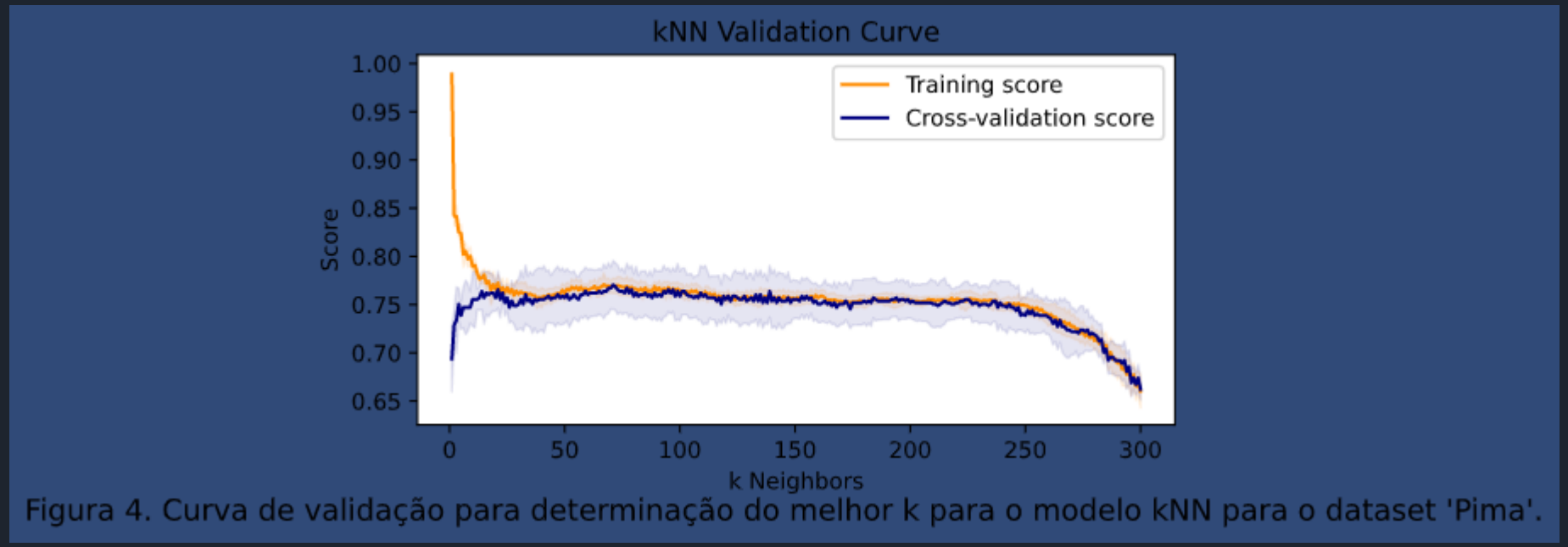

Figura 4. Curva de validação para determinação do melhor k para o modelo kNN para o dataset 'Pima'.

Questão 1 – Comparação dos modelos

- Semelhança entre resultados obtidos
- Importância das assunções dos modelos

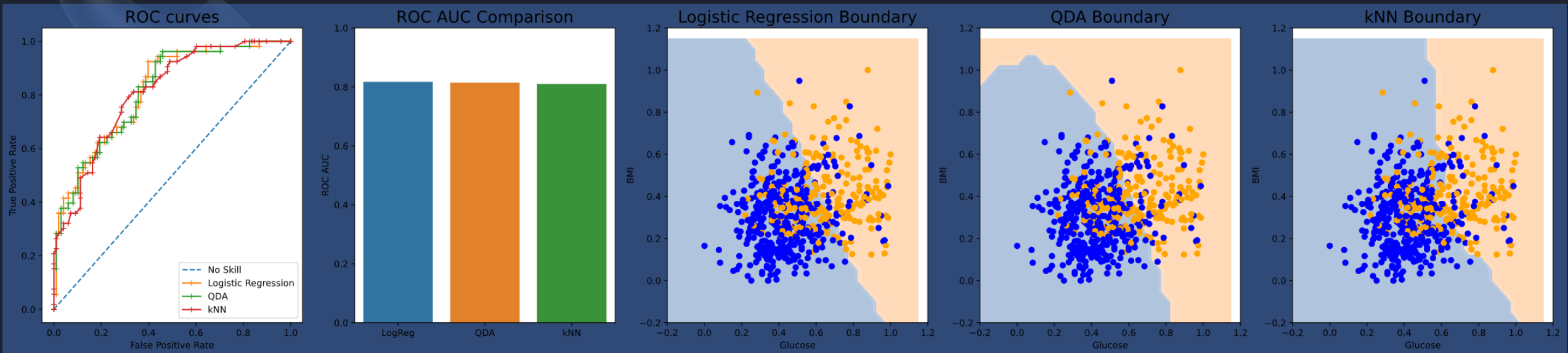

Figura 5. Curvas ROC, gráfico de barras com os valores de ROC AUC, e gráficos com as fronteiras de decisão para os três modelos testados.

Questão 2 – Regressão Logística e Árvores de Decisão

Regressão Logística

```
Weighted F1 score on training set:  0.5564
Weighted F1 score on test set:  0.5435
Weighted F1 score on 5-fold test data:  0.5409 +/- 0.027

Classification report:
              precision    recall  f1-score   support

         CYT      0.507     0.753     0.606        93
         ERL      0.000     0.000     0.000         1
         EXC      0.000     0.000     0.000         7
         ME1      0.455     0.556     0.500         9
         ME2      0.000     0.000     0.000        10
         ME3      0.667     0.688     0.677        32
         MIT      0.617     0.592     0.604        49
         NUC      0.641     0.477     0.547        86
         POX      0.667     0.500     0.571         4
         VAC      0.000     0.000     0.000         6

    accuracy                          0.569       297
   macro avg      0.355     0.356     0.351       297
weighted avg      0.541     0.569     0.544       297
```

Árvores de Decisão

```
Weighted F1 score on training set:  0.5986
Weighted F1 score on test set:  0.5512
Weighted F1 score on 5-fold test data:  0.5519 +/- 0.0611

Classification report:
              precision    recall  f1-score   support

         CYT      0.526     0.538     0.532        93
         ERL      0.000     0.000     0.000         1
         EXC      0.571     0.571     0.571         7
         ME1      0.571     0.889     0.696         9
         ME2      0.500     0.300     0.375        10
         ME3      0.711     0.844     0.771        32
         MIT      0.617     0.592     0.604        49
         NUC      0.522     0.547     0.534        86
         POX      0.000     0.000     0.000         4
         VAC      0.000     0.000     0.000         6

    accuracy                          0.566       297
   macro avg      0.402     0.428     0.408       297
weighted avg      0.542     0.566     0.551       297
```

Questão 2 - kNN

kNN (k = 17)

```
Weighted F1 score on training set:  0.6207
Weighted F1 score on test set:  0.5693
Weighted F1 score on 5-fold test data:  0.5664 +/- 0.0411

Classification report:
              precision    recall  f1-score   support

         CYT      0.530     0.656     0.587        93
         ERL      0.000     0.000     0.000         1
         EXC      0.667     0.571     0.615         7
         ME1      0.462     0.667     0.545         9
         ME2      0.333     0.200     0.250        10
         ME3      0.719     0.719     0.719        32
         MIT      0.681     0.653     0.667        49
         NUC      0.560     0.488     0.522        86
         POX      0.667     0.500     0.571         4
         VAC      0.000     0.000     0.000         6

    accuracy                          0.579       297
   macro avg      0.462     0.445     0.448       297
weighted avg      0.568     0.579     0.569       297
```

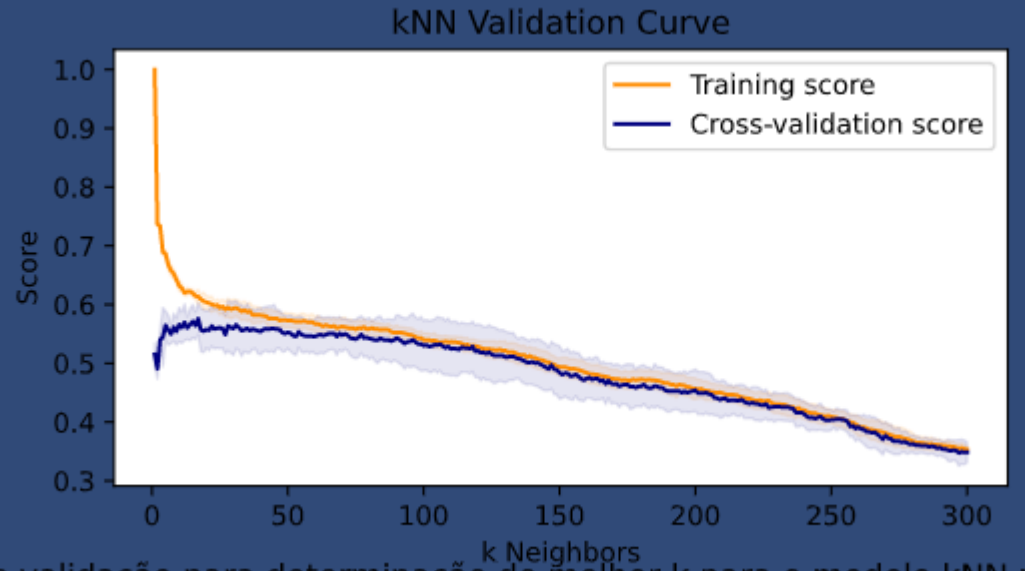

Figura 6. Curva de validação para determinação do melhor k para o modelo kNN para o dataset 'yeast'.

Questão 2 – Comparação de Resultados

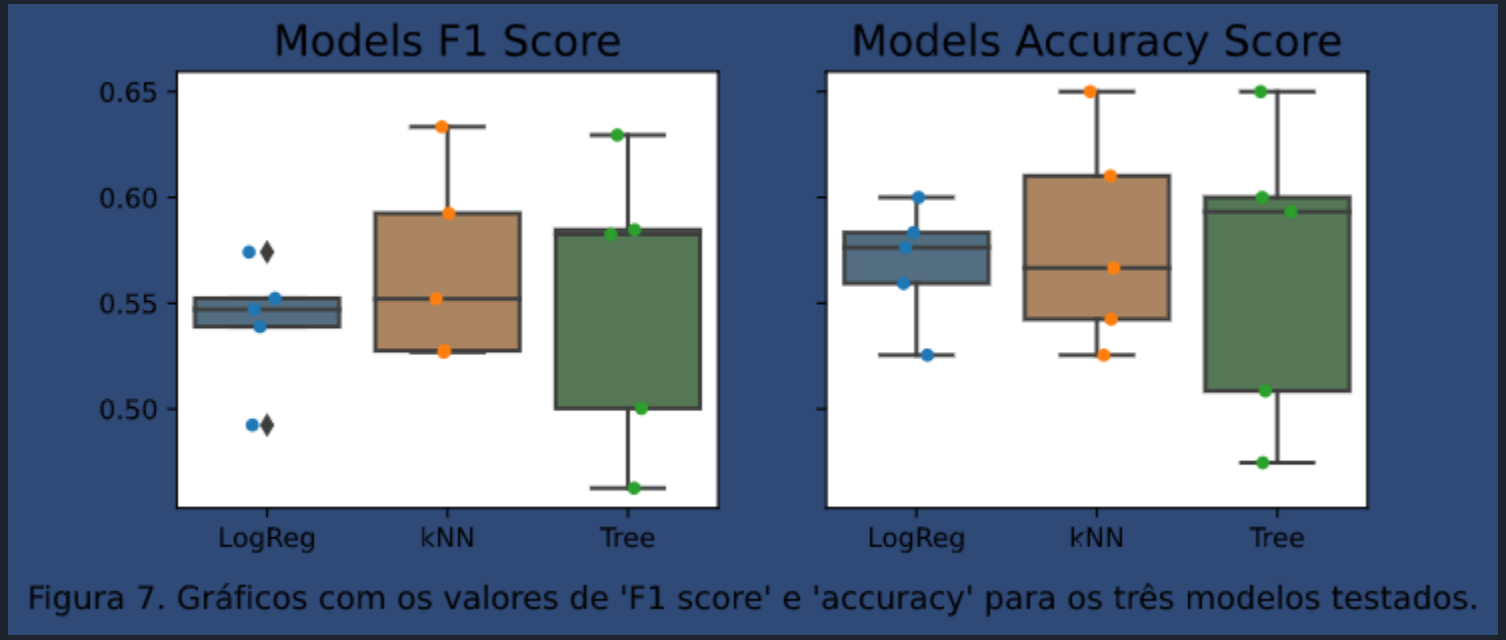

Figura 7. Gráficos com os valores de 'F1 score' e 'accuracy' para os três modelos testados.